DEVELOPMENTS IN WATER SCIENCE

45

K.W. HIPEL AND A.I. McLEOD

TIME SERIES MODELLING OF WATER RESOURCES AND ENVIRONMENTAL SYSTEMS

ELSEVIER

TIME SERIES MODELLING OF WATER RESOURCES AND ENVIRONMENTAL SYSTEMS

DEVELOPMENTS IN WATER SCIENCE, 45

OTHER TITLES IN THIS SERIES

4 J.J. FRIED
GROUNDWATER POLLUTION
13 M.A. MARIÑO AND J.N. LUTHIN
SEEPAGE AND GROUNDWATER
14 D. STEPHENSON
STORMWATER HYDROLOGY AND DRAINAGE
18 J. BALEK
HYDROLOGY AND WATER RESOURCES IN TROPICAL REGIONS
19 D. STEPHENSON
PIPEFLOW ANALYSIS
22 H.C. RIGGS
STREAMFLOW CHARACTERISTICS
25 W. KINZELBACH
GROUNDWATER MODELLING
26 D. STEPHENSON AND M.E. MEADOWS
KINEMATIC HYDROLOGY AND MODELLING
28 M. JERMAR
WATER RESOURCES AND WATER MANAGEMENT
29 G.W. ANNANDALE
RESERVOIR SEDIMENTATION
30 D. CLARKE
MICROCOMPUTER PROGRAMS IN GROUNDWATER
31 R.H. FRENCH
HYDRAULIC PROCESSES ON ALLUVIAL FANS
32 L. VOTRUBA, Z. KOS, K. NACHÁZEL, A. PATERA AND V. ZEMAN
ANALYSIS OF WATER RESOURCE SYSTEMS
33 L. VOTRUBA AND V. BROŽA
WATER MANAGEMENT IN RESERVOIRS
34 D. STEPHENSON
WATER AND WASTEWATER SYSTEMS ANALYSIS
35 M.A. CELIA ET AL. (EDITORS)
COMPUTATIONAL METHODS IN WATER RESOURCES, 1, MODELING SURFACE AND SUB-SURFACE FLOWS
36 M.A. CELIA ET AL. (EDITORS)
COMPUTATIONAL METHODS IN WATER RESOURCES, 2, NUMERICAL METHODS FOR TRANSPORT AND HYDROLOGICAL PROCESSES
37 D. CLARKE
GROUNDWATER DISCHARGE TEST SIMULATION AND ANALYSIS MICROCOMPUTER PROGRAMMES IN TURBO PASCAL
38 J. BALEK
GROUNDWATER RESOURCES ASSESSMENT
39 E. CUSTODIO AND A. GURGUI (EDITORS)
GROUNDWATER ECONOMICS
40 D. STEPHENSON
PIPELINE DESIGN FOR WATER ENGINEERS
third revised and updated edition
41 D. STEPHENSON AND M.S. PETERSON
WATER RESOURCES DEVELOPMENT IN DEVELOPING COUNTRIES
42 K. NACHÁZEL
ESTIMATION THEORY IN HYDROLOGY AND WATER SYSTEMS
43 J. ZÁRUBA
WATER HAMMER IN PIPE-LINE SYSTEMS
44 W.E. KELLY AND S. MAREŠ (EDITORS)
APPLIED GEOPHYSICS IN HYDROGEOLOGICAL

TIME SERIES MODELLING OF WATER RESOURCES AND ENVIRONMENTAL SYSTEMS

KEITH W. HIPEL

Departments of Systems Design Engineering and Statistics and Actuarial Science
University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

A. IAN McLEOD

Department of Statistical and Actuarial Sciences, The University of Western Ontario
London, Ontario, Canada, N6A 5B7
and Department of Systems Design Engineering, University of Waterloo

ELSEVIER

ELSEVIER
Amsterdam — London — New York — Tokyo 1994

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

ISBN: 0 444 89270-2

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

This book is printed on acid-free paper.

Printed in The Netherlands

To Our Wives

Sheila and Maree

and

Our Children

Melita, Lloyd, Conrad, Warren

and

Jonathan

TABLE OF CONTENTS

PREFACE

In order to understand and model how one or more inputs to a given system affect various outputs, engineers and scientists take measurements over time. For a given input or output variable that is being monitored, the set of observations appearing in chronological order is called a time series. In **time series modelling and analysis**, time series models are fitted to one or more time series describing the system for purposes which include forecasting, simulation, trend assessment, and a better understanding of the dynamics of the system. The kinds of systems which can be studied from a time series modelling viewpoint range from a purely socioeconomic system where econometricians may wish to determine how leading indicators can be used to forecast the future performance of the economy of a country, to a completely physical system for which engineers may wish to ascertain how land use changes have affected the environment.

This is a book about time series modelling of water resources and environmental systems. From the area of **stochastic hydrology**, consider for example, how time series analysis may be employed for designing and operating a system of hydroelectric facilities. After fitting stochastic or time series models to pertinent hydrological time series such as sequences of riverflows, precipitation and temperature measurements, the fitted models can be employed for simulating possible hydrological inputs to the hydroelectric system. These inputs can be used for testing the economical and physical performance of various alternative designs of the system in order to select the optimal design. Subsequent to the construction of the system of reservoirs, stochastic models can be employed for forecasting the input flows to the system and the demand for electrical consumption, in order to ascertain an optimal operating policy which maximizes the hydroelectrical output subject to physical, environmental, economical and political constraints.

As another example of the use of time series modelling in the environmental sciences, consider the use of time series models for the **trend assessment** of water quality time series. Land use changes such as increased industrialization and the cutting down of forests may cause water quality variables in a river to deteriorate over time. To model the trends and estimate their magnitudes, appropriate time series models can be employed. In Part VIII of the book, the intervention model is suggested as a flexible model for use in an **environmental impact assessment** study. From a qualitative viewpoint, the intervention model for a water quality study can be written as:

Water Quality Variable	=	**Intervention Effects**	+	**Riverflows**	+	**Other Water Quality Variables**	+	**Missing Values**	+	**Noise**

In the above relationship, the output water quality variable on the left may represent a phenomenon such as phosphorous levels in a river. The intervention effects are modelled as the changes in the mean level of the phosphorous time series due to the external interventions. The input series may consist of riverflows and other water quality variables such as temperature and turbidity. If there are not too many missing values, the intervention model can be used to estimate them. Finally, the noise term in the intervention model captures what is left over after the other model components are accounted for. Because the noise is often modelled as an **ARMA** (autoregressive-moving average) model, it can properly handle any autocorrelation which may be present. Hence, one does not have to assume that this noise is white.

The areas of **stochastic hydrology** and **statistical water quality modelling** constitute two domains that are of direct interest to scientists and engineers who wish to study water resources as well as other related environmental systems. As illustrated above, within each domain, time series modelling possesses widespread applicability. Rather than treating these areas separately, this book amalgamates these two subfields under the overall field called environmetrics. As explained in Section 1.1 of the first chapter, **environmetrics** is the development and application of statistics in the environmental sciences. Because various kinds of time series models constitute the main type of statistical tools described in this book, the title of the book is appropriately given as **"Time Series Modelling of Water Resources and Environmental Systems."** A variety of other statistical methods such as **graphical techniques**, **nonparametric trend tests** and **regression analysis** are also presented in the book.

To demonstrate how the time series models and other statistical techniques are applied in practice, **practical applications** to riverflow, water quality and other types of environmental time series are given throughout the book. However, the reader should keep in mind that the techniques can be applied to time series arising in fields falling outside the environmental areas of this book. Accordingly, the types of Professionals who may wish to use this book include:

Water Resources Engineers
Environmental Scientists
Hydrologists
Geophysicists
Geographers
Earth Scientists
Planners
Economists
Mechanical Engineers
Systems Scientists
Chemical Engineers
Management Scientists

Within each professional group, the book is designed for use by:

Teachers
Students
Researchers
Practitioners and Consultants

When employed for teaching purposes, the book can be used as a course text at the **upper undergraduate or graduate level**. Depending upon the number of topics covered, it can be utilized in a **one or two semester course**.

As can be seen from the Table of Contents, and also **Table 1.6.1**, the book is divided into **ten major Parts** having a total of **twenty-four Chapters**. For convenience, the titles of the ten Parts are listed in **Table P.1**. The book contains descriptions of **specific statistical models and methods** as well as **general methodologies** for applying the statistical techniques in practice. The only background required for understanding virtually all the material presented in the book is an introductory one semester course in probability and statistics.

Table P.1. Ten main Parts in the book.

Part Numbers	Part Titles
I	Scope and Background Material
II	Linear Nonseasonal Models
III	Model Construction
IV	Forecasting and Simulation
V	Long Memory Modelling
VI	Seasonal Models
VII	Multiple Input-Single Output Models
VIII	Intervention Analysis
IX	Multiple Input-Multiple Output Models
X	Handling Messy Environmental Data

Depending upon the background and interests of the reader, Section 1.6.2 describes the **various routes** that can be followed for exploring the countryside of ideas presented in the book. Consequently, in the Preface only the main topics covered in the book are highlighted. In general, the book progresses from describing simpler to more complicated models in order to model more complex types of environmental data sets.

A summary of the main contents of each Part in the book is presented at the start of each Part. Consequently, the reader may wish to **read each of the ten summaries** before referring to detailed descriptions of techniques and methodologies presented in each chapter. Within **Part I**, the **scope** of the book and some **basic statistical definitions** that are useful in time series modelling are given in Chapters 1 and 2, respectively. As explained in Chapter 1, statistical methods can be used to enhance the **scientific approach** to studying environmental problems which should eventually result in better overall **environmental decisions** being made at the political level of decision making. In order to give the reader some tools to work with, various classes of linear nonseasonal models are presented in **Part II**. More specifically, in Chapter 3 the **AR** (autoregressive), **MA** (moving average) and **ARMA** models are defined and some of their important theoretical properties, such as their theoretical autocorrelation structures, are derived. As is the case with all of the models defined in the book, special emphasis is placed upon highlighting **theoretical properties which are useful in practical applications**. The models of Chapter 3 are designed for application to **stationary nonseasonal time series** for which the statistical properties do not change over time. In Chapter 4, the **ARIMA** (autoregressive integrated moving average) model is defined for application to a **nonstationary nonseasonal time series** where, for instance, the level of the series may increase or decrease with time. Other kinds of time series models are presented in Parts V to IX of the book. A list of all the time series models described in the book is given in **Table 1.6.2**. However, before presenting other kinds of time series models after Part II, some practical aspects of time series modelling are described. In particular, **Part III** explains how the nonseasonal models of Part II can be fitted to yearly time series by following the identification, estimation and diagnostic check stages of **model construction**. Applications to yearly hydrological and other kinds of time series explain how this is executed in practice. The basic model building methods of Part III are simply extended for use with more complicated time series models given later in the book. Using practical applications, **Part IV** explains how the nonseasonal models of Part II can be used for **forecasting and simulation**. Forecasting and simulating with other models in this book simply involve making appropriate changes and extensions to the procedures given in Part IV.

The **Hurst Phenomenon** defined in Chapter 10 of **Part V** caused one of the most interesting and controversial debates ever to take place in hydrology. Both theoretical and empirical research related to the Hurst phenomenon are described in detail and a proper explanation for the Hurst phenomenon is put forward. One spinoff from research related to Hurst's work was the development of **long memory models** for which the theoretical autocorrelation function dies off slowly and is not summable (see Section 2.5.3 for a definition of long memory). The two types of long memory models presented in Part V are the **FGN** (Fractional Gaussian noise) model of Section 10.4 and the **FARMA** (fractional ARMA) model of Chapter 11.

The three kinds of seasonal models presented in **Part VI** are the **SARIMA** (seasonal ARIMA), **deseasonalized**, and **periodic models**. The latter two seasonal models are especially well designed for use with environmental time series in which certain kinds of stationarity are present in each season. Forecasting experiments demonstrate that **PAR** (periodic autoregressive) models provide better forecasts than their competitors when forecasting certain kinds of seasonal hydrological time series.

A major emphasis of this book is the use of time series models and other related statistical approaches in **environmental impact assessment** studies. Parts VII, VIII and X provide significant contributions to this topic. The type of **multiple input-single output model** presented in **Part VII** is the **TFN** (transfer function-noise) **model** which is designed for modelling situations qualitatively written as:

Single **Output** **Variable**	=	**Multiple** **Inputs**	+	**Noise**

In the above expression, for example, the single output variable may be riverflows which are caused by the input variables consisting of precipitation and temperature, plus an ARMA noise term. The type of basic structure contained in the TFN model reflects the physical realities present in many natural systems. Indeed, forecasting experiments described in Chapter 18 demonstrate that a TFN model provides more accurate forecasts than those obtained from a costly and complicated conceptual model.

The **intervention model** of **Part VIII** constitutes a worthwhile extension of the TFN model of **Part VII**. In addition to handling **multiple inputs** and **autocorrelated noise**, the intervention **model** has components for modelling the **effects of external interventions** upon the mean level of the output series and also for **estimating missing values**. The qualitative expression for the intervention model shown earlier in the Preface demonstrates the flexible design of the model. Indeed, extensive applications to both water quality and quantity data in Chapters 19 and 22 clearly show the great import of this model in **environmental impact assessment**.

Within **Part IX**, the class of **multiple input-multiple output models** that is described is the **multivariate ARMA** family of models. In order to reduce the number of model parameters, a special case of the multivariate ARMA models, which is called the **CARMA** (contemporaneous ARMA) set of models, is suggested for use in practical applications. Qualitatively, a multivariate ARMA model is written as:

Multiple **Outputs**	=	**Multiple** **Inputs**	+	**Noise**

This type of model is needed when there is **feedback** in the system. For instance, there can be feedback between water levels in a large lake and precipitation. Evaporation from the lake causes clouds to form and precipitation to take place. The precipitation in turn causes the lake level to rise from precipitation falling directly on the lake as well as increased riverflows into the lake from rivers affected by the precipitation.

In **Part X**, general methodologies and specific techniques are presented for **assessing trends** and other statistical characteristics that may be present in **messy environmental data**. Water quality time series, for instance, are often quite messy because there are a large number of missing observations and many outliers. To extract an optimal amount of information from messy environmental data, it is recommended to carry out both exploratory data analysis and confirmatory data analysis studies. Simple graphical methods can be used as **exploratory data analysis** tools for discovering the main statistical characteristics of the series under study. At the **confirmatory data analysis** stage, statistical models can be used to model formally the time series in order to confirm presence of the key statistical properties. After estimating missing data points, the **intervention model** is employed in Chapter 22 for modelling trends in water quality series measured in creeks that may have been influenced by cutting down a forest. Because there are a great number of missing observations for water quality variables measured in a large lake, **nonparametric trend tests** are employed in Chapter 23 for detecting any trends caused by industrialization near the lake. Finally, in Chapter 24, a general methodology is presented for detecting and analyzing trends in water quality series measured in rivers. A **robust locally weighted regression smooth** can be employed for visualizing the trend in a graph of the data. Furthermore, a flexible nonparametric trend test is used for confirming the presence of the trend. **Table 1.6.4** summarizes the **trend analysis approaches** used in the book within the overall framework of exploratory and confirmatory data analyses.

Most chapters in the book contain the following main components:

Descriptions of techniques
Representative applications
Appendices
Problems
References

Additionally, time series models presented later in the book usually constitute appropriate **expansions of the ARMA-type** models presented earlier. Finally, **flexible model construction methods** are presented for all of the classes of time series models described in the book. Consequently, the time series models are **completely operational** and can be used now within a systematic data analysis study.

Except for the long memory FGN model of Chapter 10, all of the time series models discussed in detail in this book are directly related to the basic ARMA model. Hence, the nonseasonal ARMA and ARIMA, long memory FARMA, three types of seasonal, TFN, intervention, and multivariate models of Parts II, V, VI, VII, VIII, and IX, respectively, all can be considered as belonging to the **extended family of** ARMA **models**. All of these models possess sound theoretical designs and can be conveniently applied to actual data sets using the flexible model building procedures described in the book. Furthermore, numerous practical applications and comparisons to other kinds of models clearly demonstrate the usefulness of these models in environmetrics.

Consider first the utility of the **intervention model** of Part VIII. As shown by the many applications in Part VIII and Chapter 22 of the intervention model to water quantity and quality time series, the intervention model works very well in practical applications. In fact it is probably the most useful and comprehensive time series model available for use in environmental impact assessment studies at the present time.

As summarized in **Table 1.6.3**, experimental results are provided at various locations in the book for a range of situations in which ARMA models are used for forecasting and simulation. When forecasting annual geophysical time series, **forecasting experiments** demonstrate that ARMA models and a nonparametric regression model produce more accurate forecasts than their competitors (see Section 8.3). For the case of average monthly riverflows, PAR models identified using proper identification plots provide accurate forecasts (Sections 15.3 and 15.4). As explained in Section 15.5, **combining forecasts** from different models can produce more accurate forecasts when the individual models are quite different in design and both models produce reasonably accurate forecasts. However, because SARIMA models do not forecast seasonal riverflow data nearly as well as PAR models, combining forecasts across these two models produces forecasts that are less accurate than the PAR forecasts on their own. Forecasting experiments in Chapter 18, demonstrate that a **TFN model forecasts riverflows significantly better than a conceptual hydrological model.**

Another major finding in the book is that ARMA-type models work remarkably well for **simulating** both nonseasonal and seasonal hydrological time series (see **Table 1.6.3**). As demonstrated by the **simulation experiments** outlined in Section 10.5, ARMA models statistically preserve the Hurst statistics and thereby provide a clear answer to the riddle of the Hurst phenomenon. Furthermore, as pointed out in Section 14.8, PAR models statistically preserve the critical period statistics for monthly riverflow time series.

When developing a time series model for describing a given time series, experience has shown that better models can be developed by following the identification, estimation and diagnostic check stages of **model construction.** Only a properly designed and calibrated model has the potential to work well in simulation and forecasting.

The McLeod-Hipel Time Series (MHTS) Package constitutes a flexible **decision support system** for carrying out comprehensive data analysis studies in order to obtain meaningful statistical results upon which wise decisions can be made. As explained in Section 1.7, the MHTS package can be used for fitting virtually all of the models presented in this book to sets of time series by following the three stages of **model construction.** The MHTS package can then utilize calibrated models for performing applications such as **forecasting and simulation** experiments. Moreover, the MH package is especially useful for executing **statistical environmental impact assessment** studies where a practitioner may use tools such as graphical methods, nonparametric trend tests, intervention models, and regression analysis. Part X of the book explains how these kinds of methods can be employed for retrieving useful information from **messy environmental data.**

As a closing to the Preface, the authors would like to comment upon the **future of environmetrics, in general, and time series modelling, in particular.** As world populations continue to expand, the demand for potable water as well as other natural resources will no doubt greatly increase. Certainly, more and more of the natural environment will be altered due to increased industrialization, expansion of agricultural lands and other land use changes. These man-induced activities could in turn cause a dramatic deterioration of the environment. To better understand how man's activities affect the environment, **extensive measurements** will have to be taken of a wide range of variables including water quality, water quantity and meteorological phenomena. Of course, proper **experimental design** procedures should be used for deciding upon where and when the data should be optimally collected. This vast array of observations will have to be efficiently stored in a complex **computer system** for subsequent use in **data analysis** and **decision making.** A wide range of **time series models,** including those described in this book, as well as other appropriate statistical methods, will be needed as key modelling techniques in the **scientific data analysis** studies of the huge amounts of environmental information. By properly collecting and analyzing the data, **better decisions** can be made for obtaining solutions to pressing environmental problems which minimize man's detrimental impacts upon the natural environment. Paradoxically, the future health of the environment is questionable while the futures of environmetrics and also time series modelling are indeed very promising. Certainly, environmetrics provides one of the "medicines" that can be used to help "cure" a sick patient who appears to be lapsing into a terminal illness. **The authors sincerely hope that their timely book on time series modelling of water resources and environmental systems will help to influence people for developing and adopting sound environmental policies.**

ACKNOWLEDGEMENTS

For almost two decades, the authors have been carrying out research on the development of time series models and other related statistical methods, along with their application to hydrological, water quality, as well as other kinds of environmental data sets. During this time period, the authors have had the distinct pleasure and privilege of interacting with many fine people at their own universities and many other research institutions throughout the world. Accordingly, the authors would like to take this opportunity to **express their deepest appreciation to their many colleagues, students and friends** who have contributed significantly to this book. The types of contributions range from being coauthors on research papers which form the bases of many chapters in the book to providing timely advice and encouragement during the many years over which the book was written.

For thoughtful advice, guidance and encouragement received throughout his academic career, K.W. Hipel is grateful to his friend and colleague **Prof. T.E. Unny** who passed away on December 28, 1991, before the publication of this book. A.I. McLeod is thankful for research guidance on time series modelling furnished by **Prof. M.E. Thompson**. Over the years, the authors have worked on research in time series analysis with Master's and Ph.D. students as well as other colleagues. This collaborative research and work carried out on their own, have resulted in research papers published in water resources, statistical, hydrological, environmental, operational research, and electrical engineering journals, as well as conference proceedings. Whenever material from a research publication is used in a chapter in this book, a proper reference is, of course, given. In alphabetical order, the authors' colleagues who are coauthors on research material utilized in the book are **Dr. Paul C. Baracos, Dr. Byron A. Bodo, Dr. Fernando Camacho, Dr. Francoise d'Astous, Dr. P.K. Fosu, Dr. Carlos Jimenez, Prof. William C. Lennox, Prof. Dennis P. Lettenmaier, Dr. W.K. Li, Prof. Edward A. McBean, Dr. Robert M. Thompstone, Prof. T.E. Unny, Dr. Roland R. Weiler** and **Prof. Sidney Yakowitz**. The authors would like to express their deepest gratitude to each of these authors for his or her joint research contributions and personal friendship.

Besides coauthors on papers, the authors were inspired by research produced by many other scientists who are referenced throughout the book. In many cases, the authors were most fortunate to have interesting research discussions with colleagues who provided timely information and advice. These colleagues include the late **Prof. Ven T. Chow, Prof. Jacques W. Delleur, Prof. Lucien Duckstein, Dr. Abdel El-Shaarawi, Dr. Robert M. Hirsch, Prof. Roman Krzysztofowicz, Prof. Ian B. MacNeill, Dr. N. Matalas, Prof. J.D. Salas, Prof. H.W. Shen** and **Prof. Vujica Yevjevich**.

Over the years, each author has employed earlier versions of chapters within the book in a time series modelling course at his university. Useful suggestions from **students** attending the lectures have been helpful for improving how the topics are presented in the book. For example, a former Master's student, **Mr. Michael Rooks**, suggested the final title chosen for the book, while **Ms. Larissa Stefurak**, another Master's student, as well as **Mr. Steve Fletcher** and **Mr. Kei Fukuyama**, Ph.D. candidates, put forward many useful changes.

The authors greatly appreciate the financial support furnished by the **Natural Sciences and Engineering Research Council** (NSERC) **of Canada** which funded the development of much of the authors' research presented in this book. K.W. Hipel is grateful for funding from the **Japan Society for Promotion of Science (JSPS)** and **Centre National de la Recherche Scientifique** (CNRS) in France for carrying out research during his 1984 and 1989/90 sabbatical research leaves, respectively. In addition, A.I. McLeod thanks the **Institute of Statistical Mathematics** in Tokyo for financing his research in Japan in 1990. Moreover, the **Ontario Ministry of the Environment Research Advisory Council** as well as **Environment Canada** supported research projects on trend analysis of water quality time series and stochastic hydrology, respectively. Finally, the authors appreciate the provision of research facilities at the **University of Waterloo** and the **University of Western Ontario** during the research development and writing of the book.

Ms. Norma Secord, the Administrative Assistant in the Department of Systems Design Engineering at the University of Waterloo, has done an outstanding job in expertly typing this entire book. Her excellent work and loyal service are greatly appreciated by the authors. Additionally, **Dr. Liping Fang** provided many useful suggestions regarding the technical contents and layout of the book and he helped to process the final manuscript. The authors would like to thank Dr. Fang for his fine work. Moreover, they are grateful to **Mr. Edward Volcic, Ms. Lesley Milley, Ms. Wendy Stoneman,** and **Ms. Sharon Bolender** for their assistance in preparing the camera-ready version of the book. Mr. Volcic and Ms. Milley are undergraduate students in Systems Design Engineering at the University of Waterloo and their work was partially funded by the **Faculty of Engineering** as part of their workterm experience.

Mr. Robert L. Goodman is the Director for Editorial Research in North America for Elsevier Science Publishers of New York. He has provided the authors with valuable suggestions and constant encouragement during the entire time period over which the book was written. The authors are most grateful to Bob for his thoughtful assistance.

Keith W. Hipel
Professor and *Chair*
Department of Systems Design
Engineering
Cross-Appointed Professor to
Department of
Statistics and Actuarial Sciences
University of Waterloo

A. Ian McLeod
Professor
Department of Statistical and
Actuarial Sciences
The University of Western Ontario
Adjunct Professor
Department of Systems Design Engineering
University of Waterloo

Christmas, 1993

PART 1

SCOPE

AND

BACKGROUND MATERIAL

The **objectives** of Part I of the book are to explain the important roles that time series modelling has to play in environmental decision making and to provide definitions for some basic statistical concepts that are used in time series modelling. As can be seen in the Table of Contents given at the start of the book, Part I consists of the following two chapters which are entitled:

CHAPTER 1 - ENVIRONMETRICS, SCIENCE AND DECISION MAKING

CHAPTER 2 - BASIC STATISTICAL CONCEPTS

The first chapter furnishes the basic **motivations** for writing a book on time series modelling of water resources and environmental systems as well as pointing out the import of **time series modelling in science and decision making**. Chapter 2 presents a variety of basic **statistical definitions** that are utilized in the subsequent chapters in the book.

Consider now in more detail some of the main contributions of the first two chapters, starting with Chapter 1. As explained in Section 1.1, this book on time series modelling is actually a document about **environmetrics** - the development and application of statistical methods in the environmental sciences. Because environmental data sets usually consist of observations measured over time, **time series models** constitute important statistical tools for use in environmetrics. In fact, the time series and other statistical methods presented in the book draw upon research developments from two areas of environmetrics called **stochastic hydrology** and **statistical water quality modelling** as well as research contributions from the field of **statistics**. As pointed out in Section 1.2, the use of statistical techniques can enhance the **scientific method** which in turn means that pressing environmental problems can be more efficiently and expeditiously solved. When carrying out a scientific data analysis study using environmental data such as hydrological and water quality time series, one can employ both exploratory data analysis and confirmatory data analysis tools. The purpose of **exploratory data analysis** is to use simple graphical methods to uncover the basic statistical characteristics of the data which can be modelled formally at the **confirmatory data analysis** stage utilizing time series models and other kinds of statistical methods. For example, in an **environmental impact assessment** study, exploratory graphs may clearly indicate the presence of **trends** in a water quality time series due to land use changes. The trends can then be modelled and their magnitudes estimated at the confirmatory data analysis stage using the **intervention model** of Part VIII. Section 1.3 outlines how a time series model, such as an intervention model, can be systematically fitted to a data set by following the identification, estimation and diagnostic check steps of **model construction**. By keeping in mind the basic **physical system** within which a data analysis is being carried out, one can put the overall environmental problem into proper perspective. Section 1.4 explains why the **hydrological cycle** provides a good physical structure for the types of environmental systems studied in the applications in this book. By executing a **scientific data analysis** study based upon a sound environmental system framework, one can improve environmental decision making. Section 1.5 provides a description of **engineering decision making** and explains how it can be enhanced using proper data analysis studies. Next, Section 1.6 describes the **organization** of the book and suggests various sequences of chapters that can be followed according to the needs and backgrounds of the readers. Finally, Section 1.7 describes a **decision support system,** called the **McLeod-Hipel Time Series Package,** that permits a user to take immediate advantage of the many statistical techniques presented in the book. The book should be useful for **teachers, students, researchers and practitioners** who are interested in confronting challenging data analysis problems arising in water resources and environmental engineering.

In Chapter 2, **basic statistical definitions** that are needed in time series modelling are presented. First, the different **kinds of time series** that can arise in practice are discussed. After briefly explaining what is meant by a **stochastic process**, the concepts of **stationarity** and **nonstationarity** are described. This is followed by a variety of specific statistical definitions including the **autocorrelation function** for describing linear dependence among observations in a time series. Although the time series modelling and analysis carried out in this book are mainly done in the **time domain**, some contributions from **spectral analysis** are discussed in Section 2.6.

CHAPTER 1

ENVIRONMETRICS, SCIENCE AND DECISION MAKING

1.1 THE NEW FIELD OF ENVIRONMETRICS

The **overall objectives** of this book are to present flexible statistical methodologies for scientifically carrying out data analysis studies of environmental time series and to describe a broad variety of useful statistical tools for implementing these methodologies. The methodologies include general procedures for systematically executing a data analysis study as well as the main steps required for fitting a specific statistical model to a data set. Because environmental data are almost always available as observations measured over time, most of the particular tools presented in the book consist of different kinds of time series models. However, other statistical methods such as informative graphical techniques, regression analysis and nonparametric tests are also discussed. Finally, the main types of environmental time series that are used for demonstrating how to apply the procedures and techniques consist of hydrological observations such as riverflows, precipitation and temperature series, as well as many different kinds of water quality series measured in rivers and lakes.

In fact, the contents of this book fall within a relatively new and dynamic academic discipline called **Environmetrics**. The term Environmetrics was first coined by J.S. Hunter on January 27, 1976, at a meeting of the Committee on National Statistics held at the National Academy of Sciences in Washington, D.C., and it is defined as the development and application of statistical methodologies and techniques in the environmental sciences (Hunter, 1990). The environmetrics approaches and techniques given in this book are based upon research results developed largely in the areas of statistics, stochastic hydrology and statistical water quality modelling. This book presents pertinent developments from these fields in a systematic and coherent fashion under the unifying umbrella of environmetrics. Furthermore, the title of the book reflects the fact that the time series modelling and other procedures given in the book should be especially useful for scientists, engineers and applied statisticians studying water resources and environmental systems. Nonetheless, students, teachers, practitioners, and researchers working in many other fields where time series models are applied may find much of the material to be quite helpful for addressing many different kinds of data analysis problems.

As brief illustrations of the usefulness and importance of environmetrics, consider an application from **statistical water quality modelling** and another one from stochastic hydrology. Figure 1.1.1 displays a graph of 72 average monthly phosphorous observations (in milligrams per litre) from January, 1972, until December, 1977, for measurements taken by the Ontario Ministry of the Environment downstream from the Guelph sewage treatment plant located on the Speed River in the Grand River basin, Ontario, Canada. Notice in this figure that the abscissae along the *X* axis represent time, where the monthly observations are numbered sequentially from 1 to 72. The ordinates along the *Y* axis give the values of the phosphorous concentrations in mg/l. For easier interpretation of the graph, the measurements are plotted as small circles at discrete points in time and are joined by straight lines. In February, 1974, a pollution abatement procedure was brought into effect by implementing conventional phosphorous treatment at the Guelph station. Observe in Figure 1.1.1 the manner in which the man-made intervention of phosphorous removal has dramatically decreased the mean level of the series after the intervention date. Moreover, as indicated by the blackened circles in this figure, there are missing data points both before and after the intervention date. For displaying a missing value on the graph, the missing observation is replaced by its monthly average across all of the years. However, estimating a missing monthly value by a specified monthly mean may not be an accurate procedure since the autocorrelation or dependence structure inherent in the time series is ignored and the influence of the intervention is not considered. It is explained in Chapter 19 how the **intervention model** can be used not only to estimate the missing observations where the autodependence structure among measurements is automatically taken into account but also to model statistically the effects of the tertiary phosphorous treatment for reducing the mean level of the series. In Section 19.4.5, intervention analysis is employed for realistically modelling the water quality time series of Figure 1.1.1 by constructing an appropriate intervention model. The intervention model fitted to the series in Figure 1.1.1 shows that there is a 75% drop in the mean level of the series where the 95% confidence interval is from 71% to 78%. Rigorous statistical statements like this are extremely useful in **environmental impact assessment** studies. Besides Chapter 19, other **trend analysis** procedures and applications are given in Chapters 22 to 24 of Part X.

As a second demonstration of the efficacy of environmetrics, an application from stochastic hydrology is utilized. **Stochastic hydrology** arose in the early 1960's in the field of water resources and it deals with the application of stochastic and time series models to hydrological time series (Maas et al., 1962). Because simulated sequences from time series models fitted to riverflow series are used in the design and operation of systems of reservoirs, stochastic hydrology is also referred to as **synthetic or operational hydrology.** The water quantity application involves an interesting water resources systems problem that was solved in Brazil by Silva et al. (1984). In 1984, hydroelectric plants accounted for 85% of Brazil's installed electrical capacity of 40,000 MW. In order to optimize the generation of power from a vast complex of hydroelectric plants, various types of linear multivariate time series models were used to model and simulate flows into the reservoirs (Pereira et al., 1984). To coordinate the most economical operation of both the hydrothermal generating system and the hydroelectric system, stochastic dynamic programming was used (Pereira and Pinto, 1985). Silva et al. (1984) clearly demonstrated that their use of time series models as well as other systems science techniques for optimally operating the huge electrical system clearly saved the country about $87 million U.S. in five years. Because of this great practical accomplishment, in 1985 the Institute of Management Science (TIMS) awarded the authors second prize at the 14th Annual Competition for the Edelman Award for Management Science Achievement. Time series models similar to those used in the Brazilian study are presented in Parts VII to IX in this book.

There are many specific reasons why one may employ time series models in environmental engineering. For instance, in the first application referred to above, the intervention model is utilized for quantifying the magnitude of a step trend in a water quality time series. In the

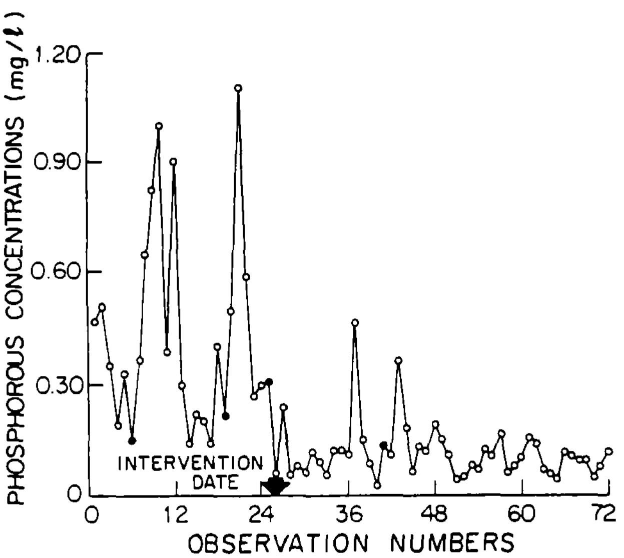


Figure 1.1.1. Monthly phosphorous data for the Speed River near Guelph, Ontario, Canada.

second application, simulated sequences and forecasts from multivariate time series models are used for maximizing profits by economically operating a complex system of hydroelectric power plants. However, there are also some very general advantages and uses for employing time series models, and, for that matter, most other types of mathematical models. Firstly, a model provides a common **communication medium** by which scientists, engineers, statisticians and other decision makers can realistically discuss an environmental problem. For instance, the graph in Figure 1.1.1 geometrically displays the obvious step drop in the phosphorous series due to the tertiary treatment. Additionally, as shown in Section 19.4, the intervention model fitted to this series accurately models this trend as well as other characteristics of the data such as autocorrelation and a pure random component. By examining the estimate for one of the parameters in the model, interested parties can see how the magnitude of the improvement in phosphorous levels is quantified. This type of representation of information and accompanying communication lead to a second important benefit of formal modelling - **understanding.** By discussing a problem with others and using mathematical models as a means of communication, the ultimate result is a better understanding of the problem by everyone concerned. In spite of the fact that the data in Figure 1.1.1 contain some randomness, missing values and a step trend, one can sort out these components using graphs and an intervention model, and, thereby, better understand what is happening. Moreover, a clearer understanding of the problem ultimately leads to **improved decision making.** Suppose, for instance, environmental authorities require that there be a 90% drop in the phosphorous levels in the river. The results of the intervention study clearly demonstrate that more intensive tertiary treatment would be required to reduce the phosphorous concentrations. Alternatively, if the environmental regulators want only an 77% reduction, one could argue that this level is almost achieved, since the 95% confidence interval for the estimated 75% reduction is from 71% to 78% and 77% is contained within this interval.

In summary, the modelling process can lead to better communication and understanding which eventually can result in improved decisions being made. However, inherent in this argument is that the mathematical model being used properly models the physical phenomena that are being studied. When modelling nature within the context of the scientific method, one should always employ a mathematical model that is realistically designed for capturing the key characteristics of the physical process being examined. One should always strive to **design a mathematical model to fit the physical problem** and never try to distort the physical process to fit a given mathematical model.

In the next section, it is explained how statistical modelling can enhance the **scientific method**. Subsequently, the main components of a **statistical scientific investigation** are pointed out and general approaches to **data analysis** are discussed. For systematically fitting a time series model to a given data set, an overall systems design approach to **model construction** is presented in Section 1.3. The **hydrological cycle** is discussed in Section 1.4 as a basic physical structure for describing the kinds of **environmental systems** studied in this book. The importance and role of environmetrics in **environmental decision making** are pointed out in Section 1.5. The **organization** of the book is thoroughly explained in Section 1.6, along with suggestions of various routes that can be followed when studying the rich variety of environmetrics techniques that are presented. Before the conclusions, a **flexible decision support system** is described in Section 1.7 for permitting a user to take full and immediate advantage of the environmetrics technologies given in the book.

1.2 THE SCIENTIFIC METHOD

1.2.1 Spaceship Earth

During the first weekend of December, 1989, the Cold War between the two superpowers came to an official end. On December 2 and 3, Soviet President Mikhail Gorbachev and American President George Bush held friendly bilateral meetings on ships anchored off the island of Malta in the Mediterranean Sea. Besides establishing a good working relationship between the two leaders, the summit's main achievement was the prospect of achieving an early agreement on decreasing by fifty per cent the superpowers' long-range nuclear arsenals. Massive nuclear and conventional weapon systems had been developed by both the Americans and Soviets during the Cold War period which lasted from the end of World War II right up until the end of the 1980's. At last, the threat of the destruction of the entire human race by a **global thermonuclear war** between the two superpowers seemed to be waning. Henceforth, the total number of nuclear weapons would subside and, hopefully, this would take place as quickly as possible.

Although the threat of extinction by a nuclear war has lessened, the citizens of the world are now well aware of an even more ubiquitous and deadly menace to human survival. This is the continuing ruination by mankind of the natural environment which supports all life forms on the planet earth. This **environmental devastation of the air, land and water** is being brought about by human activities such as cutting down forests, releasing untreated industrial and human wastes into the environment, widespread spraying of improperly tested insecticides on crops, draining too many wetlands, driving too many cars, excessive energy consumption, and, of course, overpopulation. Although proper management of the environment was not on the agenda at the Malta Summit, it seemed that nature gave Gorbachev and Bush a timely omen by flexing its muscles. On December 2, 1989, a violent storm with gusting winds of up to 100 km per hour created waves up to 4 metres high which played havoc with arrangements for the first day of the meeting. Because the Soviet missile cruiser Slava was bobbing wildly up and down like a cork in the storm, it was impossible to hold the first meeting aboard the Soviet cruiser as planned. Rather, the two Presidents were forced to meet for their first round of talks on the larger Soviet ship Maxim Gorky, which was moored in much calmer waters alongside a dock in the Maltese port of Marsaxlokk. For those who were thinking about the urgent environmental issues that these two world leaders as well as others must address, the message was clear - **nature is the key player here on earth and it should be treated with respect.**

For a dramatic example of an adverse environmental change caused by man, consider the so-called **greenhouse effect.** As reported by Levine (1990), during the past ten years, the trace gas composition of the atmosphere has been changing significantly over time. The buildup of atmospheric greenhouse gases including carbon dioxide, methane, nitrous oxide, chlorofluorocarbons and tropospheric ozone, could lead to global warming. This in turn could cause many undesirable aftermaths such as the melting of the polar icecaps and the ensuing flooding of coastal regions. Additionally, related climatic changes could turn fertile regions into deserts and thereby trigger huge migrations of populations to more hospitable regions. One major factor for the increase in CO_2 levels in the atmosphere is the conversion of forests to agricultural land through burning. Because the carbon incorporated in the trees is not balanced by carbon accumulated in crops or grasses, the burning constitutes a net release of carbon to the atmosphere. Tropical deforestation through burning is especially serious in the Amazon rainforests of Brazil. For example, during an American space shuttle flight in 1988, the astronauts photographed a biomass burn smoke cloud over the Amazon region which covered 3,000,000 km^2. The size of this cloud was second in size to the largest smoke cloud of 3,500,000 km^2 which was photographed by astronauts in 1985. Prior to 1985, the larger biomass clouds covered areas of only 300,000 km^2 (Levine, 1990). Other environmental problems caused by man-made changes to the atmosphere include depletion of stratospheric ozone which absorbs biologically lethal solar ultraviolet radiation, and acid rain which is insidiously decimating forests in the Northern Hemisphere.

The earth, in fact, has often been compared to a spaceship containing a valuable and fragile environment. It is the only known spaceship in the universe within which humans and other life-forms can live. Therefore, its natural resources should not be squandered, ruined or destroyed. As an illustration of what could happen to the entire planet, consider what may have taken place on **Easter Island** many centuries ago. Easter Island, located in the South Pacific ocean 2,700 km west of Chile, has an area of 163 km^2 and holds the distinction of being the most isolated piece of inhabited land in the world. On Easter Sunday in 1722, Jacob Roggereen, a Dutch explorer, discovered this remote island which he aptly named Easter Island. Today, the island is controlled by Chile. Easter Island is best known for its large stone statues called moai. More than 600 of these statues of grotesquely-shaped humans are scattered on the island and some of them are as high as 12 meters and weigh as much as 82 metric tons. However, most of the statues range from 3.4 to 6 meters in height. It is believed that most of these giant statues were sculpted in the period from about 1400 to 1680 A.D. What is not known is why the great moai culture that designed, built and erected these monoliths suddenly collapsed around 1680.

One hypothesis of what took place on Easter Island is provided by Flenley and King (1984) and Dransfield et al. (1984) in articles published in the journal Nature. More specifically, pollen records (Flenley and King, 1984) and shells from palm fruits coming from an extinct type of Chilean wine palm (Dransfield et al., 1984) suggest the existence of forests on the island and their decline during the last millennium. The authors of the two papers conjecture that the deforestation of Easter Island by the moai people caused their own cultural disintegration. In other words, the self-imposed environmental destruction of Easter Island led to the extinction of a great culture. Imagine what went through the minds of the remaining moai inhabitants as they cut down the last of the date palms and thereby severed their umbilical cord with nature. When one contemplates the analogy of Easter Island to the current treatment of the environment on spaceship earth by the world civilizations, the ultimate result is frightening.

The famous statistician George Box believes that the root cause of the present sorry state of the world is the **scientific method**. As explained by Box (1974), the scientific method provides the secret of learning fast and allows the normally very slow process of learning by chance experiences to be greatly accelerated. The scientific method furnished the fuel for the **industrial revolution** which started in Great Britain in the early 1700's and spread quickly to most of continental Europe and America in the 1800's. Today, the industrial revolution is a world-wide phenomenon along with the expansion into the present information age or, as it is also called, the second industrial revolution. In all of the changes brought about by the scientific method there are both advantages and drawbacks. For example, scientific medicine is responsible for fewer deaths at birth and by disease along with longer life expectancies. The disadvantage is that populations can grow too large for the environment to support properly. Scientific agricultural methods result in higher crop yields but at the expense of massive deforestation, the addition of chemical fertilizers and poisonous insecticides to the natural environment, as well as overpopulation. The scientific method furnishes the key for massive industrial expansion in order to produce great numbers of motor vehicles, lawn mowers, televisions, packaged foods and many other products that are in high demand. Unfortunately, the by-products that are endlessly dumped into the environment during the manufacture, utilization and ultimate disposal of these products are seriously polluting the air, earth and water. In short, the life support system for humanity is seriously ill because of the scientific method and it may never recover if drastic action is not taken now.

What can help to save humanity from its present dilemma? Well, mankind used the scientific method to create the current predicament and mankind can **utilize the scientific method to assist in restoring and properly managing the environment**. However, extremely quick and decisive action is required before it is too late. Box (1974) believes that **statistical methods** can act as a catalyst to further accelerate the scientific method for solving pressing environmental problems. In the next section, the scientific method is defined and the manner in which statistics can improve this powerful philosophy on the learning process is explained. Moreover, the potential influence of scientific studies of environmental problems upon the overall decision making process is described in Section 1.5. With knowledgeable and committed people at the helms of government and industry, as well as widespread public awareness, hopefully the current environmental mess can be rectified.

1.2.2 Description of the Scientific Method

Professor John Polanyi, winner of the 1986 Nobel Prize for Chemistry, presented a seminar on science, technology and society at Wilfrid Laurier University located in Waterloo, Ontario, Canada, on September 26, 1990. In his speech, he pointed out that the overall purpose of science is to **search for truth.** The general methodology which is employed to try to discover truths about nature is called the **scientific method.** Box (1974, 1976) considers the scientific method to be a process of **controlled learning**. By employing appropriate statistical methods in conjunction with the scientific method, the learning process can be made as efficient as possible. Furthermore, because this formal learning process can result in the discovery of thought-provoking and unforeseen truths, Groen et al. (1990) call science **the discipline of curiosity.**

A mandate of environmental agencies is to monitor the natural environment by taking measurements of various natural phenomena. For example, Environment Canada possesses massive statistical records on variables such as riverflows, temperature, precipitation, barometric pressures, and a wide range of water quality variables, from the east to west coasts of Canada. These huge accumulations of data on their own do not allow scientists and engineers to reach a better understanding of how various components of the environment function. On the other hand, speculative theoretical models or hypotheses about how the environment works, will not in the absence of data verification shed insight into what is taking place either. To reach a better understanding of nature through science, one must **consider both the available data and proposed theories** in order to be able to explain the behaviour of the phenomenon being studied.

Following the research of Box (1974, 1976), Figure 1.2.1 displays graphically how iterative learning between theory and data is carried out in science. In this figure, the theoretical realm of models and ideas is called **hypotheses** while the real world of facts and observations is referred to as **data**. Starting at the top left part of Figure 1.2.1, an initial hypothesis, H_1, about how nature behaves leads by a process of **deduction** to direct consequences of the theory which can be compared to the measured data. If these consequences fail to agree with the data, one can exploit the differences or errors in order to revise the hypothesis or theory by a process called **induction**. Notice in Figure 1.2.1 that induction goes from the data to the theory or, in other words, from specific facts to the general hypothesis, which is appropriately modified based upon the above stated discrepancies. Using the revised hypothesis, H_2, the learning cycle is repeated by employing deduction to go from the general to the specific. If the consequences of the new hypothesis are not in accordance with the data, one can utilize induction again as guidance for modifying the theory. These **learning cycles** consisting of deduction and induction are repeated as often as necessary until an acceptable hypothesis or theory is found. Eventually, a theory may be discovered which cannot be refuted by the available data. Further, this entire process of iterative learning leads to a much deeper understanding of what is occurring in the real world.

Figure 1.2.1 depicts the scientific learning process as an **iterative procedure**. One can also envisage the scientific method as a **feedback system**. As shown in Figure 1.2.2, the initial idea for a scientific study is stated as a hypothesis, H_1, which is then subjected to the ultimate test as to whether or not it describes what is happening in nature. More specifically, the discrepancies or errors between the data and the consequences of hypothesis H_1 lead to a modified hypothesis H_2. This feedback loop can continue so that H_2 leads to H_3 and, in general, H_i becomes H_{i+1}, until the data no longer refutes the hypothesis.

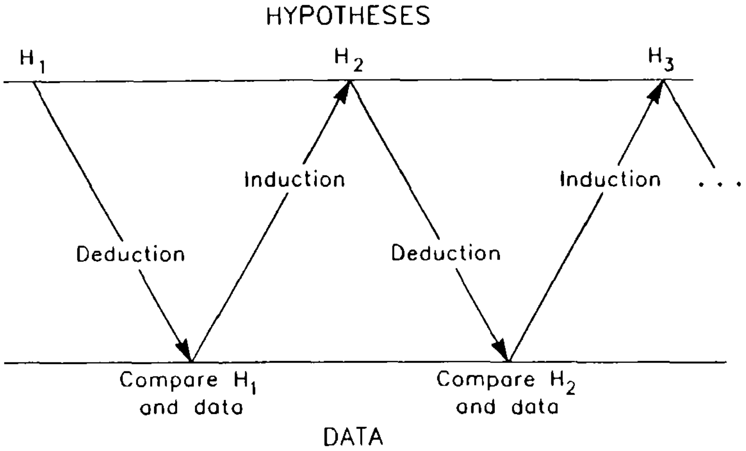


Figure 1.2.1. The iterative learning process used in science (Box, 1974, 1976).

At the bottom of Figure 1.2.2 is the **data input to the scientific method.** The data are placed here because they constitute the foundations of this whole learning procedure. As noted earlier, any hypothesis or mathematical model must reflect what is happening in the real world as represented by the observations. In the types of environmental problems used in applications in this book, the data were usually collected over a relatively long time period. For instance, average monthly riverflows are often available over a time span of 50 to 100 years. Weekly water quality data may be measured over a time period of 5 to 10 years. Whatever the case, when one is carrying out many environmental scientific studies, one can only use the available data, even though there may be many problems with these observations. There may simply not be enough time and money to obtain more data before the completion of the study. As a matter of fact, in many environmental agencies throughout the world, the scientists analyzing the natural data sets did not take part in designing the data collection procedure in the first place. Nonetheless, whenever possible, scientists are advised to assist actively in the design of the scheme for collecting the data which they will eventually analyze within the framework of the scientific method.

The methodology for efficiently collecting data for use in a scientific study is called **experimental design.** To obtain observations from the real world one must carry out experiments that are well designed. As just noted, for the case of environmental sciences such as hydrology and environmental engineering, one may have to collect data in the field over a fairly long time period. This is also the situation for areas like economics and history. However, in traditional and more basic sciences such as chemistry and physics, one can often obtain appropriate data within a fairly brief interval of time using experiments set up in the laboratory.

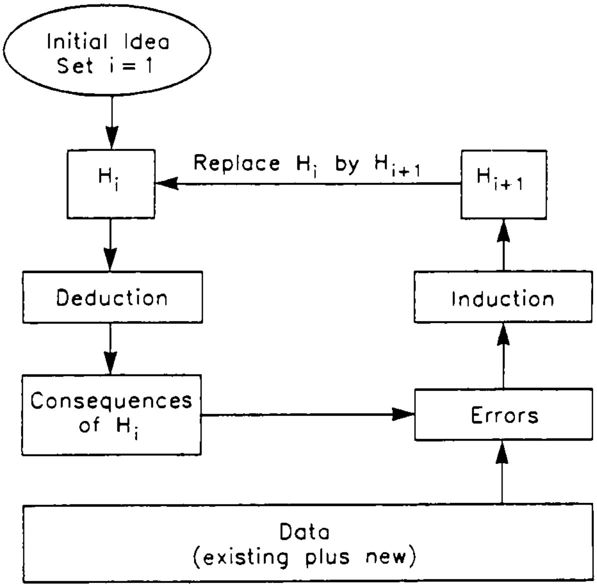


Figure 1.2.2. Feedback loop in the scientific method (Box, 1974, 1976).

To explain more specifically how experimental design plays a key role in the scientific method, refer to the expanded version of Figure 1.2.2 shown in Figure 1.2.3. At the base of the entire approach is nature depicted as a tree. An **experimental design forms a filter or window on nature** for efficiently obtaining the most appropriate observations for testing hypotheses. The available data may have been obtained from a formally designed experiment or data that were collected in an empirical fashion over the years. At each iteration in the scientific method, a current hypothesis H_i about the state of nature leads to specific consequences that are compared with facts obtained from the analysis of the available data. Differences between the consequences and facts can suggest how H_i can be modified to produce H_{i+1}. However, when it is not obvious as to what changes should be made to an unsatisfactory H_i or when further data may be required to confirm with more confidence a good hypothesis, further data should be obtained. One can see in the bottom right portion of Figure 1.2.3 that experimental design can be employed to obtain new observations. Notice that all data contain **noise** that must be taken into account in any data analysis. When experimental design is used, it is often possible to keep the noise to a minimum level, compared to when good data collection procedures are not utilized. Because the data represent the true state of nature, the scientific method leads to a convergence on the truth. If the noise level or experimental error is kept smaller, it will certainly be quicker and easier to discover the hypothesis that represents the true state of the component of nature being studied. Furthermore, even if two scientists who are separately studying the same problem start with different hypotheses and follow different routes, they will ultimately converge to the same destination when using the scientific method. For a good description of experimental design, readers may wish to refer to the textbook of Box et al. (1978) as well as references

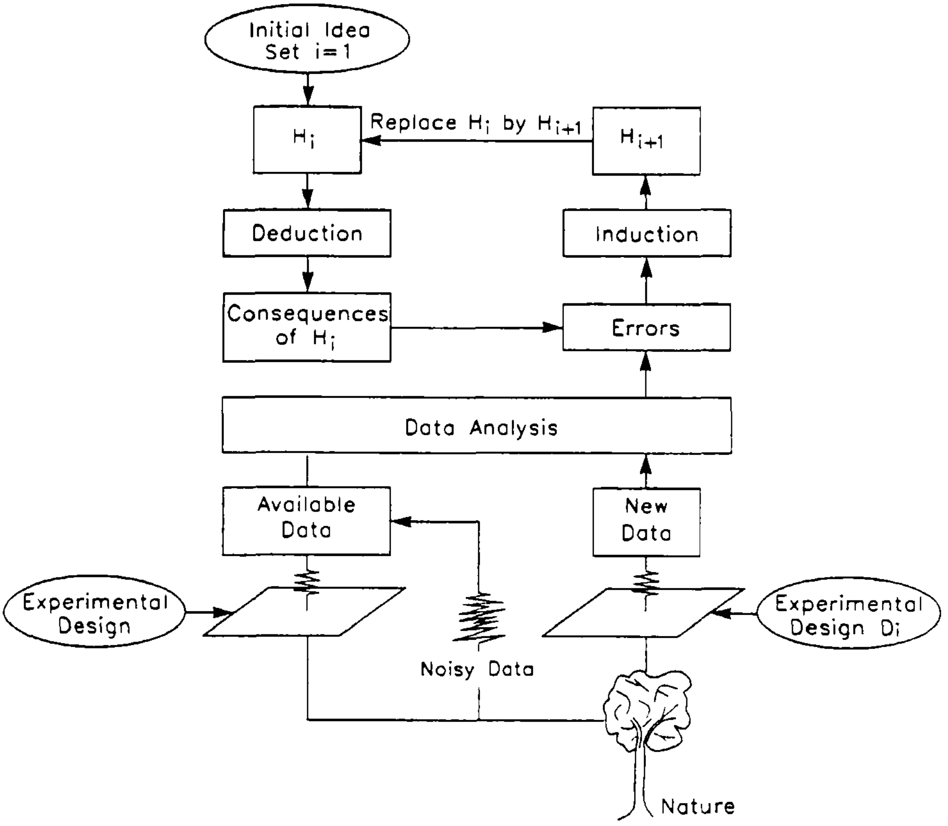


Figure 1.2.3. Data collection and analysis in the scientific method (Box, 1974, 1976; Box et al., 1978).

contained therein.

In the scientific method, one must test a hypothesis using real data. An explanation of **hypothesis testing** is given in Section 23.2 of the book just before the introduction of nonparametric statistical trend tests. The mathematical model underlying a given hypothesis is often expressed using some type of probabilistic **model**. For instance, a stochastic differential equation or a time series model may be employed to describe a given hydrological system where one might be testing a hypothesis about the output of the system given certain inputs. By definition, however, a mathematical model can never be the phenomenon it is describing, but only an approximation thereof. Nonetheless, if the mathematical model reflects well the key characteristics of the system, it may form a good basis for formally structuring the hypothesis and carrying out related data analyses. In Section 1.4.3, probabilistic models are classified according to informative criteria.

Another point to keep in mind is that the scientific method is purposefully designed to find out where one is wrong. In this way, one can **learn from experience** and come up with even a better hypothesis and underlying model (McPherson, 1990). As noted by Box (1974), there is no place in science for the man who wants to demonstrate that he has always been right.

1.2.3 Statistics in a Scientific Investigation

To carry out a **systematic scientific investigation** for discovering truths about nature, scientists and engineers employ the scientific method discussed in the previous subsection. From the outline of the main components of the scientific method, one can see that **statistics plays a key role in the scientific method** displayed in Figures 1.2.1 to 1.2.3. In fact, one can argue that by definition the scientific method must always involve statistics since one must use real data in order to refute or verify the current hypothesis. As noted by Box (1974), two main tasks in a given scientific investigation are:

1. the **design problem** where one ınust decide upon the appropriate data to obtain at each stage of an investigation.
2. the **analysis problem** where models are employed for determining what the data entitles the investigator to believe at each stage of the investigation.

In order to execute comprehensive analyses of the data, it is absolutely essential to determine properly the relevant data to obtain at the design phase by using appropriate techniques from **experimental design**. No amount of skill and experience in data analysis can extract information which is not contained in the data to begin with. Accordingly, **suitable data collection schemes are needed** for carrying out a time series analysis investigation. Within the statistical and engineering literature, extensive research has been published about designing optimal data collection schemes across a network of stations. For example, Moss (1979) wrote an introductory paper for a sequence of twenty-four papers published in Volume 15, Number 6, 1979, of Water Resources Research. For use in environmental pollution monitoring, Gilbert (1987) presents statistical methods from experimental design. Researchers at an international symposium that took place in Budapest, Hungary, delivered papers on how to design monitoring systems in order to detect changes in water quality variables (Lerner, 1986). At an international symposium held in Fort Collins, Colorado, on the design of water quality information systems (Ward et al., 1989), authors presented papers on topics ranging from data collection and network design to the roles of an information system within an overall water quality management system. Harmancioglu and Alpaslan (1992) describe water quality monitoring network design within a multiple objective framework. Other research regarding the proper design of water quality collection schemes for meeting a range of goals includes contributions by Ward et al. (1986), Ward and Loftis (1986), Whitfield (1988), and Loftis et al. (1991). Lettenmaier et al. (1978) suggest data collection schemes to use when one intends to employ the intervention model (see Section 19.7) to ascertain the effects of an intervention upon the mean level of a time series. Because most time series models must be used with a sufficient number of observations separated by equal time intervals, proper sampling is of utmost importance in time series analysis. If available measurements are not evenly spaced, appropriate **data filling** techniques can be utilized to estimate a series of evenly spaced data from the given information. As explained in Section 19.3.2, the particular technique to employ for data filling depends upon the type and amount of missing data. An advantage of employing nonparametric tests for detecting trends in time series is that they can usually be used with unequally spaced observations (see Chapter 23).

As already mentioned in Section 1.2.2, when dealing with time series studies in water resources and other environmental sciences, often the data were collected over a long period of time and the people analyzing the collected data did not take part in designing the data collection procedure in the first place. Of course, wherever possible, practitioners are advised to actively take part in the design of the scheme for collecting the data which they will analyze. Nevertheless, because environmental scientists are often confronted with analyzing data that were already collected, this book concentrates on data analysis while keeping in mind the great import of efficient data collection.

1.2.4 Data Analysis

When analyzing a given set of data within an overall scientific investigation, Tukey (1977) recommends adhering to the following two steps:

1. **exploratory data analysis,**
2. **confirmatory data analysis.**

The main purpose of the exploratory data analysis phase of data analysis is to discover important statistical properties in the given observations by carrying out simple graphical and numerical studies. The major objective of the confirmatory data analysis stage is to confirm statistically in a rigorous fashion the absence or presence of certain properties in the data.

In Part X of the book, it is explained how useful exploratory and confirmatory data analysis tools can be effectively employed for studying environmental data. Hydrological time series, such as seasonal riverflows, temperature and precipitation, are usually quite suitable for analysis purposes since, for example, they possess few missing values and outliers. However, other types of environmental series like water quality time series are often quite **messy** due to many factors. For instance, water quality time series may possess many missing observations among which there are long periods of time for which there are no measurements. Moreover, the data may have many extreme values and be affected by external interventions such as industrial development and other land use changes in a river basin. Fortunately, the **data analysis procedures of Part X are designed to handle both well behaved and messy time series.**

Many of the exploratory data analysis tools are presented in Part X, although **graphical procedures** are used throughout the book for explanation purposes. A wealth of **time series models** that can be used in data analysis studies are presented in Parts II to IX in the book. Additionally, **nonparametric trend tests** and **regression analysis methods** are discussed in Chapters 23 and 24, respectively. These latter two types of confirmatory data analysis techniques are especially well designed for use with messy environmental data.

The graph in Figure 1.1.1 of the average monthly phosphorous data demonstrates how a useful exploratory data analysis tool can usually convey a wealth of information. For instance, as already pointed out in Section 1.1, one can clearly see the drop in the mean level of the series due to the introduction of phosphorous treatment. Moreover, the location of the missing values marked as filled-in circles can be clearly seen. Within a scientific study, one may wish to test the hypothesis that there is a significant step drop in the mean level of the phosphorous and also to estimate its magnitude. The **intervention model** of Part VIII in the book can be employed for these purposes. More specifically, the details of this trend analysis assessment for the phosphorous data are presented in Section 19.4.5. From a qualitative viewpoint, the intervention model for the phosphorous series possesses the components shown below:

Phosphorous Series	=	**Intervention Component**	+	**Missing Value Terms**	+	**Correlated Noise**

Notice that this flexible model can simultaneously model the effect of the intervention, estimate the four missing values, and handle a correlated noise term. The estimate of the parameter in the intervention component, along with its standard error of estimation, allows one to carry out a hypothesis or significance test to see if there is a drop in the mean level and to quantify the magnitude of the drop. Indeed, as would be expected from the results of the exploratory graph in Figure 1.1.1, there is a significant step trend in the mean level. The best estimate for this drop is a 76% decrease compared to the previous mean, where the 95% confidence interval spans from 68% to 84%.

The exploratory and confirmatory data analysis study for the phosphorous time series points out a major contribution of this book - **the use of statistical methods in environmental impact assessment.** In a statistical environmental impact assessment, one is often required to detect and model trends in time series. In Section 22.3, a variety of **useful graphical techniques** are presented that can be employed as exploratory data analysis tools for confirming the presence of suspected trends as well as discovering unknown trends. Additionally, in Section 24.2.2, a regression analysis technique is described for tracing trends on the graph of a data set. The confirmatory data analysis tools that can be employed in trend assessment are:

1. **intervention analysis** (Chapter 19 and Section 22.4),
2. **nonparametric tests** (Chapter 23),
3. **regression analysis** (Chapter 24).

Additionally, **overall approaches for analyzing messy environmental data** that fall under the paradigm of exploratory and confirmatory data analysis are presented in Chapter 22, Section 23.5 and Section 24.3. In all three cases, detailed environmental applications are given to explain clearly how the methodologies work.

Each of the case studies in statistical environmental impact assessment presented in the book deal with the modelling and analysis of trends caused by one or more external interventions that have already taken place. For example, the step decrease in the monthly phosphorous level in the time series shown in Figure 1.1.1 from January, 1972 to December, 1977, was created by the tertiary phosphorous treatment which started in February, 1974. Consequently, the formal trend analysis of the phosphorous observations constitutes a **posterior environmental impact assessment.** In some situations, it may be required to find out the potential effects upon the environment of a planned project, before permission is granted for commencing construction. For instance, when designing a series of reservoirs for the production of hydroelectric power and other benefits, decision makers may wish to know the potential impacts of the scheme upon a range of hydrological and other environmental variables. Hence, scientists would perform an a **priori environmental impact assessment** using appropriate scientific tools. Although this book does not consider a priori environmental impact assessments, some of the statistical tools used in the many posterior environmental impact studies that are presented could provide some guidance in a priori studies. For example, knowledge gained from intervention investigations of reservoirs that have already been built could be employed for simulating possible environmental scenarios for planned reservoir systems.

1.3 PHILOSOPHY OF MODEL BUILDING

1.3.1 Occam's Razor

To better understand and control his environment, **mankind uses models.** In order to be sufficiently accurate and realistic, a model must be able to capture mathematically the key characteristics of a system being studied. At the same time, a model must be designed in a simple straightforward manner so that it can be easily understood, manipulated and interpreted. Because of the great complexity of water resources and other natural systems, models are extensively developed and applied in water resources and environmental engineering.

Time series models constitute an important class of models which can be used for addressing a wide range of challenging problems in the environmental sciences. In fact, **time series models** are a type of stochastic model designed for fitting to observations available at discrete points in time. An example of a time series is the graph of the average monthly phosphorous measurements for the Speed River shown in Figure 1.1.1. The specific type of time series model that is fitted to this series in Section 19.4.5 is the intervention model, briefly referred to in Section 1.2.4.

As explained in the previous subsection, a comprehensive **data analysis** study can be carried out by following the two main steps consisting of exploratory and confirmatory data analysis. Subsequent to employing informative graphical methods at the exploratory data analysis stage to appreciate the main statistical properties of the observations being studied, formal modelling can be done at the confirmatory data analysis stage to ascertain more precisely how the data are behaving. For instance, one may wish to use an intervention model to test the hypothesis that there is a step drop in the mean level of the series in Figure 1.1.1 and to estimate the magnitude of this decrease (see Section 19.4.5).

In the next subsection, a systematic procedure is outlined for determining the most appropriate time series model to fit to a given data set. This general model building approach is, in fact, adhered to for applying all of the time series models presented in this book to measurements taken over time. The basic idea underlying the **model construction** procedure of Section 1.3.2 is to identify a simple model which has as few model parameters as possible in order to provide a good statistical fit to the data.

The **principle of model parsimony** has historical roots that go back far into the past. Aristotle, for example, postulated that nature operates in the shortest possible way. A 14*th* century English Franciscan monk by the name of William of Occam (1280-1349) developed a principle now known as **Occam's razor**. One version of his principle states that when faced with competing explanations choose the most simple one. This is analogous to using a sharp thin razor to make a clean cut through some material. Another equivalent statement for Occam's razor is entities are not to be multiplied without necessity. Bertrand Russel (1946), the famous 20*th* century British mathematician, found Occam's razor to be very informative in logical analysis. This is because there is only one explanation or description of something which is minimum, whereas there can be an infinity of explanations which bring in other entities. Russell went on to claim that adherence to the minimum necessary explanation or description ensures that the examination of hypotheses and evidence for and against them will remain coherent. Checkland (1981) provides a good explanation of Occam's principle in his book on the theory and practice of systems engineering.

In fact, model parsimony is a key assumption embedded in the scientific method of Section 1.2. One should always strive to model nature and postulate hypotheses thereof in the most straightforward and simple manner, while still maintaining an accurate description of the phenomenon being modelled. Further discussions on modelling philosophies in time series analysis are presented in Section 5.2 as well as Section 6.3.

1.3.2 Model Construction

In time series modelling and analysis, one wishes to determine the most appropriate stochastic or time series model to fit to a given data set at the confirmatory data analysis stage. No matter what type of stochastic model is to be fitted to a given data set, it is recommended to follow the identification, estimation, and diagnostic check stages of model construction (Box and Jenkins, 1976). At the **identification stage**, the more appropriate models to fit to the data can be tentatively selected by examining various types of graphs. Some of the identification information may already be available from studies completed at the exploratory data analysis step discussed in Section 1.2.4. Because there may be a range of different families of stochastic models which can be fitted to the time series under consideration, one must choose the one or more families of models which are the most suitable to consider. The family selections can be based upon a sound physical understanding of the problem, output from the identification stage, and exploratory data analyses. Although sometimes it is possible to choose the best model from one or more families based solely upon identification results, in practice it is often not obvious which model is most appropriate and hence two or three models must be tentatively entertained. At the **estimation stage**, maximum likelihood estimates can be obtained for the model parameters and subsequently the fitted model can be subjected to **diagnostic checks** to ensure that the key modelling assumptions are satisfied. When considering linear stochastic models such as the ARMA (autoregressive-moving average) models of Chapter 3, one should check that the model residuals are not correlated, possess constant variance (i.e., homoscedasticity) and are approximately normally distributed. If the residuals are not white noise, the model should be redesigned by repeating the three phases of model construction. In practice, it has been found that a suitable Box-Cox power transformation (Box and Cox, 1964) (see Section 3.4.5) can rectify anomalies such as heteroscedasticity and non-normality. The specific tools utilized at the three stages of model construction are dependent upon the particular family of models being entertained. The logic underlying the traditional approach to model construction is displayed as a flowchart in Figure 1.3.1.

In Part III of this book, **a wide variety of model building tools** are presented for use with ARMA (defined in Chapter 3) and ARIMA (autoregressive integrated moving average) (Chapter 4) models that can be fitted to nonseasonal stationary and nonstationarity time series, respectively. Most of the identification, estimation and diagnostic check methods described in Chapters 5, 6, and 7, respectively, in Part III, can be expanded for employment with the many other kinds of time series models presented in this book.

1.3.3 Automatic Selection Criteria

As noted in Section 1.3.1, a basic tenet of model building is to keep the model as simple as possible but at the same time provide a good fit to the data being modelled. **Automatic selection criteria** (ASC) are now available for balancing the apparently contradictory goals of good statistical fit and model simplicity. In Section 6.3, a number of ASC are defined and it is

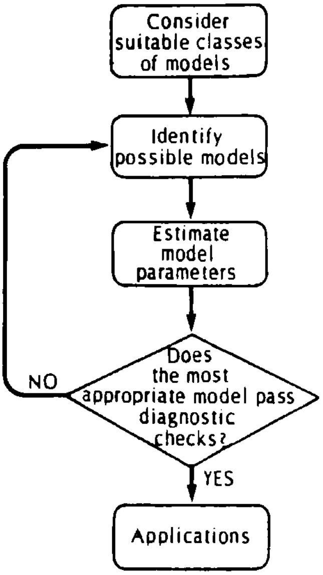

Consider suitable classes of models
Identify possible models
Estimate model parameters
Does the most appropriate model pass diagnostic checks?
NO
YES
Applications

Figure 1.3.1. Model construction.

explained how they can enhance the three stages of model construction portrayed in Figure 1.3.1. One example of an ASC is the **Akaike information criterion** (AIC) of Akaike (1974) which is used throughout this book.

In general, an ASC is defined as follows:

$$\text{ASC} = \text{Good Statistical Fit} + \text{Complexity}$$

The first term on the right hand side is written as some function of the value of the maximized likelihood function for the model fitted to the data (see Section 6.2 for a discussion of maximum likelihood estimation). This term is defined in such a way that the smaller the value the better the statistical fit. One would expect that a more complex model would furnish a more accurate description of the data. The purpose of the second entry in the ASC formula is to guard against having a model which is too complex and to abide by the principle of Occam's razor of Section 1.3.1. The complexity component in the ASC is a function of the number of model parameters, where a smaller value means that there are fewer parameters in the model. Hence, overall one would like to select the model which has the lowest value for the ASC. In Chapter 6, Figure 6.3.1 depicts how an ASC can be used in model construction.

1.4 THE HYDROLOGICAL CYCLE

1.4.1 Environmental Systems

The basic structure of the scientific method described in Section 1.2 is depicted in Figure 1.2.3. Because the scientific method seeks to discover truths about nature, the natural world as represented by the tree in Figure 1.2.3 is shown as the foundation of the scientific method at the bottom of the figure. In order for the scientific method to work, one must **employ mathematical models that reflect the important physical characteristics of the system being studied.**

When carrying out a scientific study, one can better appreciate what is taking place if one envisions a conceptual framework of the physical world within which the mathematical modelling is being done. An ideal paradigm for accomplishing this is the systems design approach to modelling. In **systems modelling**, one thinks of an overall system composed of subsystems that interact with one another in some sort of hierarchical manner. Individual subsystems can be studied and analyzed in detail using powerful mathematical models. By properly connecting the subsystems together, one can synthesize the overall system behaviour, given the initial conditions and operating rules within the subsystems.

A question that naturally arises is how one should define the boundaries of the system and its subsystems given the type of problem being studied. As explained by authors such as White et al. (1984) and Bennett and Chorley (1978), there are many ways in which one can define **environmental systems**. At the solar system level, the system consists of the sun and each of the nine planets which rotate around the sun. The yearly rotation of the earth around the sun along with the tilting of the earth's axis is the main cause of the seasons on earth. The rotation of the moon about the earth creates tides in the oceans and seas. Besides gravitational forces, direct solar energy from the sun constitutes the major input for environmental systems contained on, in and around each planet. If one is studying the overall environmental system for the planet earth on its own, then one must subdivide this system into finer subsystems which may not be explicitly considered when looking at the earth and the other planets at the solar system level. More specifically, at this level one may wish to subdivide the overall global environmental system into a number of major subsystems which include the atmosphere, hydrosphere and lithosphere (earth's crust). The biosphere subsystem, in which plant and animal life exist, occurs at the transition zone between the lithosphere and atmosphere as well as within the hydrosphere. Different kinds of ecosystems are contained within the biosphere system.

Any of the aforementioned global subsystems can, of course, be further subdivided into finer subsystems. For instance, the atmosphere can be vertically categorized from the earth's surface outward, into subsystems consisting of the troposphere, stratosphere, mesosphere and thermosphere. Within the lower part of the atmosphere, one can examine the various circulation subsystems around the world.

In summary, one can **define environmental systems in a hierarchical fashion** from overall large systems at the planetary level to very detailed subsystems at much lower levels. The system definitions to be entertained depend upon the particular problem being studied. For example, if one is examining water pollution problems within the Great Lakes in North America, the largest environmental system to consider may be the drainage basin for the Great Lakes. Within this overall system, one could examine river subsystems through which pollutants carried by water can flow into the Great Lakes. The definitions of subsystems could be made as detailed as required for the problem at hand. Within and among the subsystems, appropriate mathematical models could be used to model precisely the physical, chemical and biological interactions.

One particular environmental system which may form a basis for many problems examined by water resources and environmental engineers is the well known **hydrological cycle.** This important environmental system is described in the next subsection. Subsequently, a range of mathematical models that can be used for modelling natural phenomena within the hydrological cycle, as well as other environmental systems, are classified according to informative criteria. Based on these discussions, one can appreciate the role of time series models for describing and analyzing important environmental phenomena.

1.4.2 Description of the Hydrological Cycle

Hydrology is the science of water. In particular, hydrology deals with the distribution and circulation of water on the surface of the land, underground and in the atmosphere. Additionally, hydrology is concerned with the physical and chemical properties of water and its relationships to living things. Similar definitions for hydrology can be found in hydrological books written by authors such as Eagleson (1970), Linsley et al. (1982) and McCuen (1989).

The environmental system which hydrologists employ to describe the components of their science is called the **hydrological cycle.** Figure 1.4.1 displays a schematic of the hydrological cycle which is based upon the figure provided by Eagleson (1970, p. 6). The **throughput** to the hydrological system in Figure 1.4.1 is water which can occur in a liquid, solid or vapour phase. Because the hydrological cycle does not allow water to escape, it forms a **closed system** with respect to water. The **main forces** which propel the water through the hydrological cycle are solar energy and gravity. As pointed out by Eagleson (1970, p. 5), the dynamic processes of vapour formation and transport are powered by solar energy while precipitation formation and the flow of liquid water are driven by gravity. The transformation of water from one phase to another as well as the transportation of water from one physical location to another are the main features of the hydrological cycle. To understand the possible routes and phases a water molecule passes through in the hydrological cycle in Figure 1.4.1, one can start at any point in the cycle. Note that all water is returned from the atmosphere to land or surface bodies of water through the process of **precipitation** by going from vapour to liquid form. Liquid water can **infiltrate** into the soil and flow overland via streams, rivers and lakes to the oceans. **Evaporation** is the dynamic process which returns water from its liquid phase in streams, rivers, lakes and oceans to its vapour phase in the atmosphere. Water molecules can also be released to the atmosphere from the surfaces of plants through a process called **transpiration. Evapotranspiration** consists of the total water transferred to the atmosphere by transpiration from plants plus evaporation from the soil on which the plants are growing. It is interesting to note that in land areas having a temperate climate, approximately 70% of the precipitation returns to the atmosphere via evapotranspiration while the remaining 30% mainly appears as riverflows (McCuen, 1989, p. 668). **Sublimation** takes place when water is transformed from its solid form directly to its gaseous state in the atmosphere. Notice also in Figure 1.4.1 the various ways in which water can enter and leave the two subsystems consisting of soil and underground aquifers.

Interest in ideas related to hydrology can be traced back to the ancient Egyptians, Greeks and Romans. For instance, as early as 3,000 B.C., the Egyptians had gauges called nilometers to measure the stages or depths of the Nile River. A **nilometer** consists of a stone pillar on which markings indicate the depth of the Nile, especially during flooding. However, it was not until

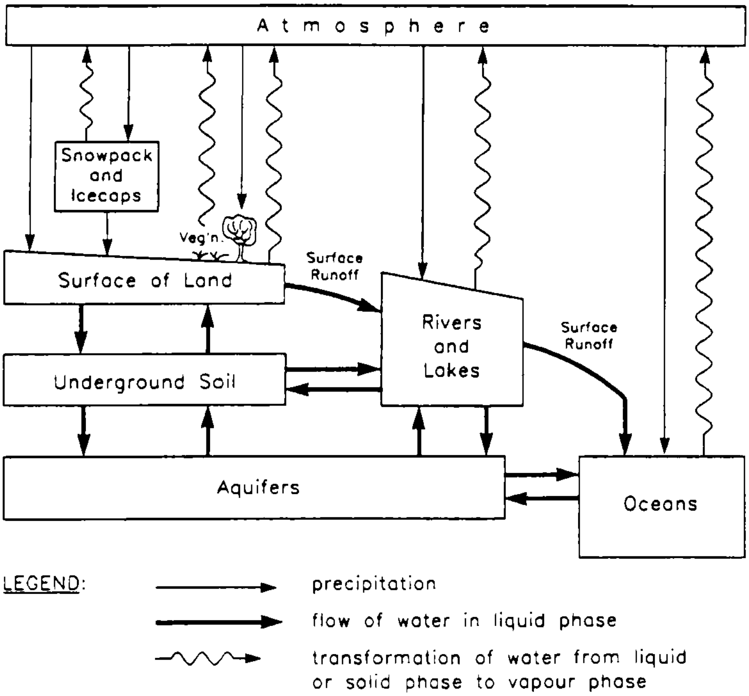


Figure 1.4.1. The hydrological cycle (Eagleson, 1970, p. 6).

the **Renaissance** that the first essentially correct picture of the hydrological cycle was produced by **Leonardo da Vinci.** Hence, this first accurate description of an important environmental system is now more than 400 years old. Besides providing a good conceptual description about how the transformation and transportation of quantities of water takes place on earth, the hydrologic cycle can be used for other purposes. Specifically, the hydrologic cycle furnishes a framework for understanding how man-made pollution can enter the hydrologic system at any point and pollute our entire environment by following the ancient pathways traced out by water in all of its forms. **Therefore, the science of hydrology provides solid foundations upon which many other environmental sciences can build and interact.**

During the past century, the development of hydrology has been mainly led by civil and agricultural engineers working on traditional engineering problems such as water supply and flood control. Consequently, the field of hydrology has been pragmatic in its outlook and narrowly focused upon only a few aspects of the overall hydrological cycle displayed in Figure 1.4.1. **However, in order to make wise and timely decisions regarding solutions to pressing environmental problems occurring throughout the hydrologic cycle and right up to the global level, a fundamental scientific understanding of hydrology is required.** Accordingly, in 1987 the National Research Council of the United States established a panel to conduct an assessment of hydrology that appeared as a report (National Research Council, 1991) which is summarized by the panel chairperson P.S. Eagleson (1991). The authors of this report, as well as Falkenmark (1990), describe many key research areas in which the science of hydrology should be expanded so that sound environmental policies can be properly devised and implemented. Indeed, they correctly point out that hydrology should be developed into a comprehensive and distinct geoscience. Moreover, they note that human activity and decision making can have dramatic influences upon the hydrological cycle and hence are an integral part of that cycle.

1.4.3 Classifying Mathematical Models

A general framework in which to envision how various components from nature interact with one another is to employ the environmental systems approach outlined in Section 1.4.1. A particularly informative environmental system for use as a sound scientific structure in water resources and environmental engineering is the hydrological cycle of Section 1.4.2 and Figure 1.4.1. Within a given environmental systems framework such as the hydrological cycle, one can use the scientific method of Section 1.2 to develop a range of specific models to describe natural phenomena occurring in the system. In order to abide by the principle of Occam's razor of Section 1.3.1, the simplest models to describe nature are almost always formulated in terms of appropriate mathematical equations. Hence, one can argue that **mathematics is the language of science**. Indeed, some mathematicians feel that mathematics should be considered as a separate scientific discipline. In reality, mathematics constitutes an interdisciplinary scientific field which supports all areas of science.

One should keep in mind that **real problems in the natural and social world inspired scientists to develop the most useful and dramatic contributions to scientific mathematics.** For example, to describe properly his revolutionary laws of nature, Sir Issac Newton developed the mathematics of calculus. In order to solve practical problems in agriculture, Sir Ronald A. Fisher invented experimental design and many other branches of statistics (Box, J., 1978). Worthwhile areas of game theory were formulated for modelling and analyzing actual social disputes (see discussion in Section 1.5.2).

Since the time of Newton, great progress has been accomplished in building a treasure house of many different kinds of mathematical models. To appreciate the general types of mathematical models that are available for application to scientific problems, it is informative to classify these models according to useful criteria. Two overall categories into which mathematical models can be placed are deterministic and stochastic models. When a mathematical model can be employed for determining exactly all the states of a system, the model is said to be **deterministic**. For instance, when plotting on a graph all values of an algebraic function representing the states of a system, the precise locations of all possible points on the curve are known because the equation is deterministic.

If a state of a system can only be described using probabilistic statements and hence its precise value is not known, the mathematical equations describing the system are said to be **stochastic** or **probabilistic**. For example, when forecasting tomorrow's weather conditions using an appropriate stochastic model along with the latest meteorological information, a meteorologist may state that there is an 80% chance of snowfall tomorrow. This means that there is still a 20% chance that no precipitation may occur. The meteorologist may go on to forecast that he or she is 95% confident that the amount of snow accumulation will be between 15 and 20 cm.

Most natural phenomena occurring in environmental systems appear to behave in random or probabilistic ways. In other words, it is almost impossible to say exactly how nature will behave in the future, although one can often make reasonable predictions about occurrences using probabilistic statements. Consequently, the mathematical tools presented in this book fall within the realm of stochastic models.

As explained in Section 2.3, **stochastic models** are mathematical models for describing systems which evolve over time according to probabilistic laws. When discussing the theoretical aspects of stochastic models, the term stochastic processes is often used in place of stochastic models. Following Cox and Miller (1965), Table 1.4.1 describes a method for categorizing stochastic models according to the two criteria of time and state space. Notice that **time** can be either discrete or continuous. The **state space** or values of the variables describing the system being modelled can also be subdivided according to discrete and continuous values. Examples of the four kind of models that can be categorized using the above criteria are given in Table 1.4.1. Markov chains, for instance, fall under the subdivision of stochastic models which incorporate discrete time and discrete values of the state space in their mathematical structure. Stochastic differential equations can handle continuous time and continuous values of the state space (Kloeden and Platen, 1992). Point processes, such as Poisson processes, model discrete values over continuous time.

Table 1.4.1. Classifications of stochastic models.

		STATE SPACE	
		Discrete	Continuous
TIME	Discrete	Markov Chains	Time Series Models
	Continuous	Point Processes	Stochastic Differential Equations

This book deals with stochastic models that model continuous observations measured at discrete points in time. Because these models formally describe measurements available over discrete time in the form of a time series, they are usually referred to as **time series models**. The application of time series models to actual data is popularly referred to as **time series analysis**.

Why are time series models of such great import in the environmental sciences? The answer is quite simple. **In order to understand how a natural system is behaving, scientists take measurements over time,** hopefully according to a proper experimental design (see Section 1.2.3). An example of a water quality time series is displayed graphically in Figure 1.1.1. Time series models are specifically designed for formally modelling this type of information which occurs frequently in practice. Furthermore, techniques for fitting time series models to data are now highly developed (see Sections 1.3.2 and 1.3.3) and, hence, these time series models can be immediately employed for modelling, analyzing and better understanding pressing scientific problems.

In 1970, Box and Jenkins dramatically launched time series modelling into the realm of real world applications with the publication of their seminal book entitled "Time Series Analysis: Forecasting and Control" (the second edition was published in 1976). Besides presenting a wide variety of useful time series models, they showed how these models can be applied to practical problems in a wide range of disciplines. One should keep in mind that both **Box and Jenkins considered themselves foremost to be scientists** and not mathematicians. In other words, these scientists developed and used mathematical tools for scientifically studying actual problems.

Some of the time series models presented in this book are also discussed by Box and Jenkins (1976). However, **in this book the latest developments in time series modelling are given,** including many new procedures for allowing models to be conveniently fitted to data within the framework of model construction outlined in Section 1.3.2. Moreover, this book presents other kinds of models, such as many of those given in Parts V to X, which are especially useful in water resources and environmental engineering. As mentioned in Section 1.1, this book is a document falling within the challenging field of environmetrics and contains a range of useful time series models which are currently completely operational.

When using a mathematical model to describe a natural system, one would like the model to have a **sound physical basis,** in addition to possessing attractive mathematical properties. As pointed out in various sections of this book, certain time series models are well designed for various kinds of applications in hydrology and water quality modelling. For example, as explained in Section 3.5, ARMA models possess good physical justifications for use in modelling annual streamflows (Salas and Smith, 1981).

1.5 DECISION MAKING

1.5.1 Engineering Decision Making

The **scientific method** of Section 1.2.2 provides a solid foundation upon which solutions to the physical aspects of environmental problems can be properly designed and tested. As pointed out in Section 1.2.3, **statistical and stochastic models** have a key role to play for enhancing scientific investigations in terms of accuracy, speed and better understanding. Furthermore, by keeping in mind the **hydrological cycle** of Figure 1.4.1, the overall physical relevance of environmental problems being studied can be kept in correct perspective.

The **physical characteristics of environmental investigations** can involve physical (ex. waterflow), chemical and biological factors. For instance, one may wish to examine how industrial chemical pollutants discharged into rivers affect certain populations of fish and suggest correct measures for overcoming any serious problems. Scientists may discover why and how fish populations are dwindling and be able to design pollution controls to rectify the situation. However, to implement corrective measures, finances are required and political decisions must be made. Therefore, in addition to generating physical solutions to a given environmental problem by use of the scientific method, scientists must also take into account the **socio-economic aspects of decision making.** In other words, **both the physical realm of nature as well as the social world created by mankind's ability to think, must be properly accounted for in real world decision making.**

As an example of a planned large-scale engineering project which could adversely affect the environment consider the case of the **Garrison Diversion Unit** (GDU). As explained by Hipel and Fraser (1980) and Fraser and Hipel (1984), the GDU is a partially constructed multipurpose water resources project in the United States which involves the transfer of water from the Missouri River basin to areas in central and eastern North Dakota that are mainly located within the Hudson Bay drainage basin. Figure 1.5.1, which is taken from Hipel and Fraser (1980) and also Fraser and Hipel (1984, p. 26), shows the major regions affected by this large project located in the geographical centre of the North American continent. After the system becomes operational, water will be pumped from Lake Sakakawea on the Missouri River via the McClusky canal to the Lonetree Reservoir located in the Hudson Bay drainage basin. From the Lonetree Reservoir water will flow along the Velva and New Rockford canals to major irrigation areas. Additionally, water from the Lonetree Reservoir will augment flow in the James River for downstream irrigation. The resulting runoff from the irrigated fields would flow via the Red and Souris Rivers into the Canadian province of Manitoba. **Adverse environmental effects** from the GDU include high pollution levels of the irrigation waters, increased chances of flooding in the Souris River, and the possibility of catastrophic environmental damage caused by foreign biota from the Missouri River basin destroying indigenous biota, such as certain fish species, in the Hudson Bay drainage basin.

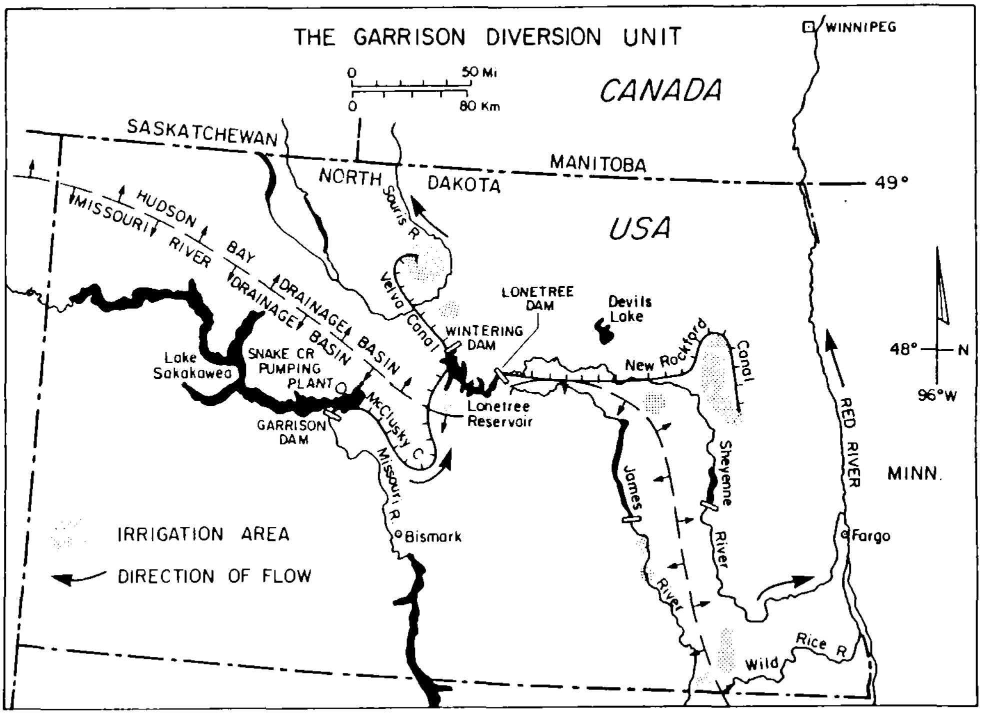


Figure 1.5.1. Map of the Garrison Diversion Unit (Hipel and Fraser, 1980; Fraser and Hipel, 1984, p. 26).

Because the GDU involved a **variety of interest groups** which interpreted the problem from different perspectives, the GDU project escalated into a **serious international environmental conflict.** The American proponents of the GDU wanted the full project, as approved by the U.S. Congress in 1965, to be built. However, Canada was afraid of potentially disastrous environmental consequences within its own borders while American environmentalists did not like some adverse environmental effects which could take place in North Dakota. The GDU controversy also involved the International Joint Commission (IJC), an impartial body initiated by the Boundary Waters Treaty of 1909 between the U.S. and Canada for investigating conflicts arising over water quantity and quality.

In order for a large-scale project like the GDU to be eventually implemented and brought into operation, the following factors must be adequately satisfied:

1. **Proper Physical Design** - For instance, physical structures such as dams, pumping stations and irrigation channels for the GDU must be correctly designed so that natural physical laws are not violated and the project is safe.
2. **Environmentally Sound Project** - If, for example, the GDU project were to be built and come into operation, adverse environmental consequences must be less than agreed upon levels. The ability to meet environmental standards must be incorporated into the basic design of the project.
3. **Economical and Financial Viability** - For the case of the GDU, the project must be economically feasible and sufficient financial resources must be available to pay for the project. It is interesting to note that some benefit-cost ratios for the GDU produced ratios much less than one.
4. **Socially and Politically Feasibility** - A politically feasible solution to the GDU project must be found before it can come into operation.

Unfortunately, for the case of the GDU only the first factor of the four listed above was ever properly satisfied. The Garrison dam on the Missouri River (see Figure 1.5.1) was completed by the Bureau of Reclamation of the U.S. Department of the Interior in 1955. Other physical facilities, such as the canals shown in Figure 1.5.1, were designed but never completely constructed. Environmental effects of the project were almost entirely ignored in the initial design of the project. **The social repercussions caused by the ensuing political controversy over environmental problems as well as suspect economic studies eventually prevented the completion of the project,** even though hundreds of millions of dollars had already been spent.

The main lesson garnered from this GDU fiasco is that all of the factors given above must be properly taken into account by scientists and other decision makers in any engineering project. Otherwise, the project may never be completed as first envisioned or it may be cancelled altogether. The scientific method and related mathematical modelling are especially important for ensuring **physical and environmental soundness.** Indeed, one should also follow a scientific approach in socio-economic modelling and analyses. Because the last two factors given above involve activities that relate directly to mankind as distinct from his natural environment, the models for studying these activities are sometimes referred to as **decision making tools.** As explained in Section 1.5.2, many of these decision making methods were developed within a field called **operational research.**

Figure 1.5.2 summarizes a **systems design approach to decision making** in engineering, which must properly consider the important factors mentioned before. To keep in mind that the entire activity takes place within the environment or natural world, the flowchart is enclosed by a wavy line. Notice that the physical, environmental, economical and financial considerations provide background information that can affect the preferences and actions of the decision makers who are included in the social and political modelling and analyses. If, for example, a first class environmental impact assessment is carried out beforehand for the project, there is a higher probability that decision makers will approve a design which abides by the suggested environmental standards. Additionally, if this project is economically and financially viable, the project has an even higher chance of being accepted by the decision makers. The political game that could take place among the decision makers who can influence the final decision can be modelled using techniques from **conflict analysis** (Fang et al., 1993; Hipel, 1990; Fraser and Hipel, 1984). The results of all of these formal studies provide background information upon which the actual decision makers can base their decisions. As shown by the feedback loops in Figure 1.5.2, additional information can be obtained as required and appropriate changes can be made. Moreover, some decision makers may obtain some of their information directly from their own observations of real world events. Hence, in Figure 1.5.2 there is an arrow going from the real world to the box labelled information for decision makers to indicate that people do not have to rely entirely upon results generated from formal studies. Finally, the design problem referred to in Figure 1.5.2 is not restricted to the construction of a new project such as the GDU. It can also represent situations such as a change in operating policy of a system of reservoir and the design of pollution control devices for installation in existing industrial facilities.

The remainder of Section 1.5, deals mainly with mathematical models that can be employed for modelling and analyzing decision making. In the next subsection, decision making models from the field of operational research are classified according to useful criteria. Subsequently, the use of conflict analysis for modelling and analyzing the GDU dispute is described and the importance of **sound scientific modelling within the overall decision making process** is once again emphasized.

1.5.2 Decision Making Techniques in Operational Research

The field of **operational research** consists of some general methodologies and many specific techniques for studying decision making problems. The British initiated operational research just prior to World War II when they performed **research** studies into the **operational** aspects of radar systems for detecting incoming enemy aircraft to the United Kingdom. Throughout the war, the British employed OR in all of their military services for successfully solving large scale military problems involving the movement of great numbers of military personnel and huge quantities of war materials (Blackett, 1962; Waddington, 1973). The American military also used this systems science approach to problem solving during the second world war but called it **operations research**. Many practitioners now simply refer to operational research or operations research as OR. Since the war, OR has been extensively expanded and utilized for looking at operational problems in many different fields outside of the military such as management sciences, transportation engineering, water resources and industrial engineering. Operational research societies have sprung up in most industrialized countries along with the publication of many OR journals.

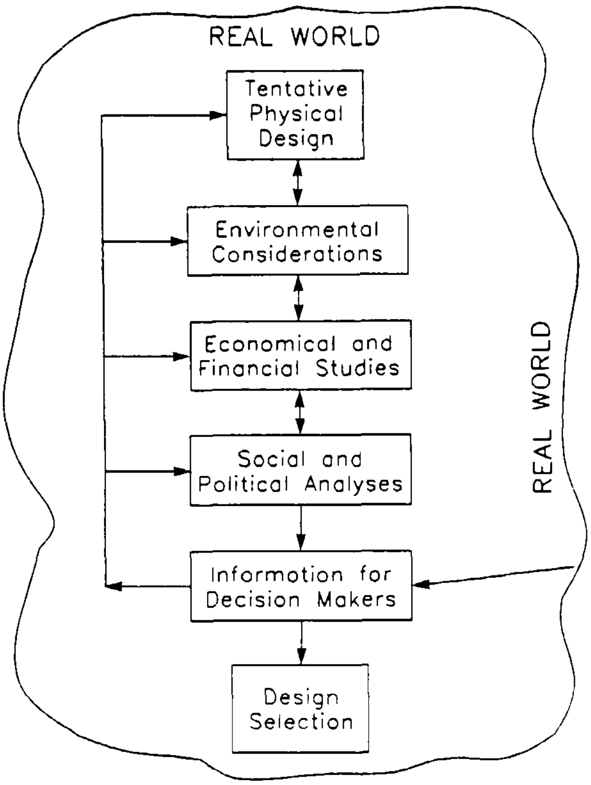

REAL WORLD
Tentative Physical Design
Environmental Considerations
Economical and Financial Studies
Social and Political Analyses
Information for Decision Makers
Design Selection
REAL WORLD

Figure 1.5.2. Engineering decision making.

The discipline of OR **is both an art and a craft**. The art encompasses general approaches for solving complex operational problems while the craft consists of a great variety of mathematical techniques which are meant to furnish reasonable results when properly applied to specific problems. Following Hipel (1990), Table 1.5.1 shows how **OR methods can be categorized according to the criteria of number of decision makers and number of objectives.** As shown in that table, most OR techniques reflect the viewpoint of one decision maker having one objective. Optimization techniques including linear and nonlinear programming, fall under this category because usually they are employed for minimizing costs in terms of dollars or maximizing monetary benefits from one group's viewpoint subject to various constraints. Often both economical and physical constraints can be incorporated into the constraint equations in optimization problems. Many of the probabilistic techniques like queueing theory, inventory theory, decision theory and Markov chains, fall under the top left cell in Table 1.5.1. An example of a technique designed for handling multiple objectives for a single decision maker is multicriterion modelling (see, for instance, Vincke (1992), Radford (1989), Roy (1985) and Goicoechea et al. (1982)). This method is designed for finding the more preferred alternative solutions to a problem where discrete alternatives are evaluated against criteria ranging from cost (a quantitative criterion) to aesthetics (a qualitative criterion). The evaluations of the criteria for each alternative reflect the objectives or preferences of the single decision maker. In Table 1.5.1, team theory is categorized according to multiple decision makers and one objective because in a sporting event, for instance, each team has the single objective of winning.

Conflict analysis (Fraser and Hipel, 1984), as well as an improved version of conflict analysis called the **graph model for conflict resolution** (Fang et al., 1993), constitute examples of techniques in Table 1.5.1 which can be used for modelling and analyzing disputes in which there are two or more decision makers, each of whom can have multiple objectives. Conflict analysis is a branch of game theory which was specifically designed and developed for studying problems in **multiple objective-multiple participant decision making.** In fact, conflict analysis constitutes a significant expansion of metagame analysis (Howard, 1971) which in turn is radically different from classical game theory (von Neumann and Morgenstern, 1953). A comparison of game theory techniques, their usefulness in OR as well as a description of present and possible future developments are provided by Hipel (1990). Additionally, Hipel (1990) explains how appropriate OR methods can be employed for studying both tactical and strategic problems which arise in decision making. In particular, conflict analysis is especially well designed for handling decision making at the strategic level where often compromise solutions must be reached in order to satisfy a wide range of different interest groups. Within the next subsection, the GDU environmental conflict introduced in Section 1.5.1 is employed to demonstrate how conflict analysis and other techniques like statistical methods from environmetrics can be used for systematically studying complex environmental problems.

Table 1.5.1 Classifications of decision making techniques.

		OBJECTIVES	
		One	Two or More
DECISION MAKERS	One	Most OR Methods	Multicriterion Modelling
	Two or More	Team Theory	Conflict Analysis

Operational research is probably the largest and most widely **known field** within which formal decision making techniques have been developed. Nonetheless, since World War II, other **systems sciences** fields have been started for efficiently solving well structured problems in order to satisfy specified objectives. Besides OR, the systems sciences include **systems engineering** (see, for instance, Checkland (1981) and references contained therein) and **systems analysis** (Miser and Quade, 1985, 1988). A large number of standard textbooks on OR are now available, such as contributions by Hillier and Lieberman (1990) and Wagner (1975). An interesting book on the application of OR methods to various water resources problems, is provided by Loucks et al. (1981). Unfortunately, none of the classical OR texts satisfactorily deal with problems having multiple decision makers (i.e., the second row in Table 1.5.1). However, Rosenhead (1989) and Hipel (1990) clearly point out directions in which OR and other systems sciences fields should be expanded so that more complex problems having many decision makers, unclear objectives and other difficult characteristics, can be properly modelled. Indeed, fields outside of OR, such as **artificial intelligence and expert systems**, as well as **information and decision technologies**, are already tackling challenging research problems in decision making at the strategic level where situations are usually not well structured. A monograph edited by Hipel (1992) on **multiple objective decision making in water resources** contains a sequence of eighteen papers regarding some of the latest developments in tactical and strategic OR techniques along with their application to challenging water resources systems problems.

1.5.3 Conflict Analysis of the Garrison Diversion Unit Dispute

The GDU dispute is employed in this section to explain how the engineering decision making procedure of Figure 1.5.2 can be carried out in practice. As pointed out in Section 1.5.1, the GDU is an important environmental dispute between Canada and the United States. Figure 1.5.1 depicts the location of the GDU irrigation scheme along with the physical facilities such as dams, canals and irrigation fields.

Hipel and Fraser (1980) and Fraser and Hipel (1984) carried out a metagame analysis and conflict analysis, respectively, of the GDU conflict as it existed in 1976, while Fang et al. (1988, 1993) also performed an analysis of the dispute for the situation that took place in 1984. To explain the type of engineering decision making used in the GDU conflict, only the results for 1976 are utilized.

Figure 1.5.3 depicts the general procedure for applying conflict analysis to an actual dispute. Initially, the real world conflict may seem to be confusing and difficult to comprehend. Nonetheless, by systematically applying conflict analysis according to the two main stages of **modelling** and **analysis** the controversy can be better understood in terms of its essential characteristics and potential resolutions. The **modelling stage** consists of ascertaining all the **decision makers** as well as each decision maker's **options** and **relative preferences**. At the **analysis stage**, one can calculate the **stability** of every possible **state** (also called an outcome or scenario) from each decision maker's viewpoint. A state is stable for a decision maker if it is not advantageous for the decision maker to move unilaterally away from it. **Equilibria** or compromise resolutions are states that are stable for every decision maker. The results of the stability analysis can be studied and interpreted by actual decision makers or other interested parties in order to understand their meaning in terms of the actual conflict. These findings may suggest types of **sensitivity analyses** that can be carried out, for example, by seeing how appropriate preference changes affect the overall equilibria. Moreover, the feedback arrows in Figure 1.5.3 indicate that the procedure for applying conflict analysis is done in an **iterative fashion**.

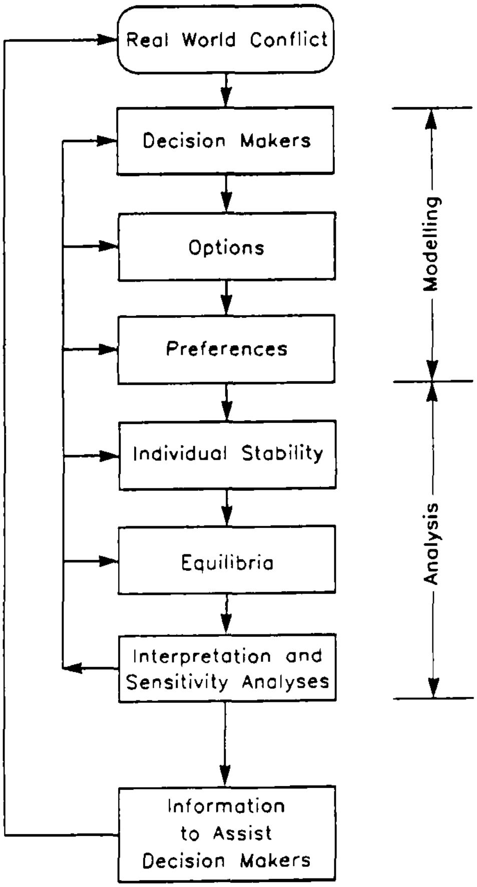


Figure 1.5.3. Applying conflict analysis.

To explain briefly the modelling and analysis of the GDU conflict using the approach of Figure 1.5.3, consider the dispute as it existed in 1976. Table 1.5.2 lists the **decision makers** involved in the conflict along with the courses of actions or **options** available to each decision maker. Briefly, the **U.S. Support** for the project consists of the U.S. Bureau of Reclamation of the U.S. Department of the Interior, the State of North Dakota and support groups within North Dakota. As shown in Table 1.5.2, the U.S. support has three mutually exclusive options available to it. The first one is to proceed to complete the GDU project as approved by Congress while the other two options constitute reduced versions of the full project to appease the Canadian Opposition (option 2) or the U.S. Opposition (option 3).

The **U.S. Opposition** consists mainly of environmental organizations such as the National Audubon Society and the Environmental Protection Agency. It has the single option of taking legal action against the project based upon American environmental legislation (option 4).

Table 1.5.2. Conflict model for the GDU dispute.

Decision Makers and Options	Representative State	
U.S. Support		
1. Complete full GDU.	N	Strategy for U.S. Support
2. Building GDU modified to reduce Canadian impacts.	N	
3. Construct GDU modified to appease U.S..	Y	
U.S. Opposition		
4. Legal action based on environmental legislation.	N	Strategy for U.S. Opposition
Canadian Opposition		
5. Legal action based on the Boundary Waters Treaty of 1909.	Y	Strategy for Canadian Opposition
International Joint Commission (IJC)		
6. Support full GDU.	N	
7. Recommend GDU modified to reduce Canadian impacts.	N	Strategy for IJC
8. Support suspension of GDU except for the Lonetree Reservoir.	Y	
9. Recommend cancellation of the GDU.	N	

The main **Canadian organizations opposed to the GDU** are the Federal Government in Ottawa, the Manitoba Provincial Government and Canadian environmental groups. The single course of action available to the Canadian Opposition is the ability to take legal action based upon the Boundary Waters Treaty of 1909 between the United States and Canada (option 5). This treaty confers legal rights to Canadians citizens for taking action in American courts when water quantity or quality is infringed upon by the Americans. The Americans also have the same rights under this treaty for entering Canadian courts.

The **International Joint Commission** (IJC) was formed under Article VI of the Boundary Waters Treaty as an impartial body to investigate water and other disputes arising between the two nations. Three Canadians and three Americans form the IJC. Whenever the IJC is called upon by the U.S. and Canada to look at a problem, it employs the best scientists from both countries to carry out rigorous scientific investigations in order to come up with a proper environmental impact assessment. As a matter of fact, the quality **of the work of the IJC is so well respected that its findings usually significantly influence the preferences of decision makers** involved in a given dispute. Because the GDU project is concerned with water quality, the IJC can only give a recommended solution to the problem. As shown in Table 1.5.2, the IJC has basically four mutually exclusive recommendations it could make after its study is completed. One option is to support completion of the full GDU project (option 6) while the other three (options 7 to 9) are reduced versions thereof.

Given the decision makers and options, one can determine strategies for each decision maker and overall states. A **strategy** is formed when a decision maker decides which of his or her options to select and which ones he or she will not take. To explain this further, refer to the column of Y's and N's in Table 1.5.2. A "Y" means that "yes" the option opposite the Y is taken by the decision maker controlling it while a "N" indicates "no" the option is not selected. Notice from Table 1.5.2, that the strategy taken by the U.S. Support is where it takes option 3 and not options 1 and 2. Possible strategies for each of the other three decision makers are also indicated in Table 1.5.2.

After each decision maker selects a strategy, a **state** is formed. Writing horizontally in text the vertical state listed in Table 1.5.2, state (NNY N Y NNYN) is created by the U.S. Support, U.S. Opposition, Canadian Opposition and IJC choosing strategies (NNY), (N), (Y) and (NNYN), respectively, in order to form the overall state.

In the GDU conflict, there is a total of 9 options. Because each option can be either selected or rejected, there is a total of 2^9=512 possible states. However, many of these states cannot take place in the actual conflict because they are **infeasible** for a variety of reasons. For instance, options numbered 1 to 3 are mutually exclusive for the U.S. Support because it can only build one alternative project. Hence, any state which contains a strategy in which the U.S. selects more than one option is infeasible. Likewise, options 6 to 9 are mutually exclusive for the IJC. When all of the infeasible outcomes are removed from the game, less than 50 out of 512 outcomes are left.

From Figure 1.5.3, the final step in modelling a conflict is to obtain relative **preferences** for each decision maker. The most precise type of preference information needed in a conflict study is **ordinal** where states are ranked from most to least preferred. This could include sets of states for which states are equally preferred within each set. Conflict analysis can also handle more general types of preferences such as **intransitive** preferences where a decision maker prefers state x to y, y to z, but z to x. When preferences are **transitive**, a decision maker prefers state x to y, y to z, and x to z. Cardinal utility functions are not used to represent preferences in conflict analysis because in practice they are almost impossible to obtain.

Preferences are often directly expressed in terms of options. For example, the **U.S. Support** most prefers states in which it selects its first option while the IJC chooses option 6. Assuming transitivity, algorithms are available to transform option preferences to state preferences such that the states are ranked from most to least preferred.

To give a general idea of the preferences in the GDU dispute, a preference description for the U.S. Support is now continued. Compared to states for which the full GDU is built, the U.S. Support prefers less states where the full GDU project is not built and the IJC also recommends this.

The **U.S. Opposition** prefers to take legal action if the full project were built. On the other hand, the U.S. Opposition would prefer not to press legal action if the U.S. were to select its third option.

The **Canadian Opposition** most prefers that no project be built. If option 2 were not chosen by the U.S. Support, the Canadian Opposition would prefer to take legal action based on the Boundary Waters Treaty. However, if the IJC recommended a given alternative which the U.S. Support decided to follow, the Canadian Opposition would prefer not to oppose it by going to court. This is because the IJC always carries out first class scientific and economic studies which are greatly respected by the Canadian government as well as others. Additionally, if the dispute were to end up in an American court or perhaps at the international court in the Hague, the court would probably follow the recommendations of the IJC in its ruling. Consequently, this application clearly demonstrates how **good science can dramatically affect the strategic decision making.** In Figure 1.5.2 proper scientific studies are used for the basic design of the project as well as the environmental considerations. Sections 1.2 to 1.4 outline how scientific investigations and related mathematical modelling can be carried out in practice.

Because the **IJC** is an impartial body, all of the states are equally preferred for it prior to the release of its comprehensive International Garrison Diversion Study Board Report in October, 1976. The report, which was commissioned by the IJC, consists of an overall report plus five detailed reports given as appendices. The five appendices are entitled Water Quality, Water Quantity, Biology, Uses, and Engineering Reports. All of the reports, and especially the first two given above, make extensive use of statistics. Indeed, **the GDU study constitutes an excellent example of how an environmental impact assessment should be executed.**

The GDU conflict model is now fully calibrated in terms of decision makers, their options and their preferences. This game model provides the basic structure within which the **possible strategic interactions** among the decision makers can be studied. More specifically, the systematic examination of the possible moves and counter moves by the decision makers during the possible evolutions of the conflict and the calculation of the most likely resolutions are referred to as the **stability analysis** stage (see Figure 1.5.3). The results of a stability analysis, including sensitivity analyses, can be used, for example, to help support decisions made by people having real power in a conflict. In practice, one would, of course, use a **decision support system** to carry out all the calculations and provide requested advice to a decision maker.

The details of the stability calculations are not given here but can be found in Chapter 2 of the book by Fraser and Hipel (1984) as well as Chapter 6 in the text of Fang et al. (1993). The stability analysis for the GDU was carried out for the situation that existed just prior to the release of the study board reports for the IJC in October, 1976. The state given in Table 1.5.2 is one of the **equilibria** predicted by the conflict analysis study and the one that occurred historically. Notice in Table 1.5.2 that the U.S. Support is going to build a project to appease the American environmentalists and, hence, the U.S. Opposition is not going to court. However, the Canadian Opposition will go to court under the Boundary Waters Treaty because the IJC is recommending a reduced version of the project (option 8).

In the **1984 conflict analysis study**, further reductions were made to the GDU project (Fang et al., 1988). As a matter of fact, all portions of the project that could adversely affect Canada were cancelled due to increased political pressures.

The GDU conflict vividly emphasizes the importance of following the main steps of the engineering decision making procedure of Figure 1.5.2. **Competent scientific and economic studies of the project by the IJC affected directly the political decision making taking place at the strategic level.** For instance, the Canadian Opposition and other interested groups put great faith in the IJC study board reports and this in turn influenced their strategic behaviour in 1976, especially in terms of preferences. By 1984, others were also influenced by **good science** and this led directly to the cancellation of most of this irrigation project which was clearly shown to be environmentally unsound.

This book deals exclusively with environmetrics. However, one should always keep in mind how the results of environmetric and other related scientific studies fit into the overall decision making process of Figure 1.5.2. Even though it may sometimes take a long time, **good science can have highly beneficial effects on decision making.**

1.6 ORGANIZATION OF THE BOOK

1.6.1 The Audience

As defined in Section 1.1, **environmetrics** is the development and application of statistical methods in the environmental sciences. This book focuses upon useful developments in environmetrics coming from the fields of **statistics, stochastic hydrology and statistical water quality modelling.** Of particular interest are time series models that can be employed in the design and operation of large-scale water resources projects, as well as time series models, regression analysis methods and nonparametric methods that can be used in trend assessments of water quality time series. In other words, this book deals with the statistical analyses of both water quantity and water quality problems. Moreover, it clearly explains how these problems can be jointly considered when carrying out environmental impact assessment studies involving trend assessment of water quality variables under the influence of riverflows, seasonality and other complicating factors.

Who will wish to study, apply and perhaps further develop the environmetrics technologies presented in this book? For sure, the environmetrics techniques should be of direct interest to **teachers, students, practitioners and researchers working in water resources and environmental engineering.** However, people from other fields who often consider environmental issues, should also find the contents of this book to be beneficial for systematically investigating their environmental problems. For example, **geographers, civil engineers, urban planners, agricultural engineers, landscape architects** and many others may wish to apply environmetrics methods given in this book to specific environmental problems that arise in their professions. Moreover, keeping in mind the great import of environmetrics and other scientific approaches in the overall decision making process (see Section 1.5), there may be many other professionals who may wish to better understand environmetrics and use it to improve their decision making capabilities. This group of professionals includes **management scientists, operational research workers, business administrators** and **lawyers** who may be employed by government agencies or industry. Finally, because most of the time series models can also be applied to data that are not environmental, there are other professionals who may find this book to be a valuable reference. **Economists**, for instance, who apply time series models to different kinds of economic time series may find useful results in this book that can assist them in their field of study. As a matter of fact, econometrics is defined as the development and application of statistical methods in economics. **Chemical engineers** may also discover useful ideas in the book for application to chemical processes involving input-output relationships.

From a **teaching viewpoint**, this book is designed for use as a course text at the upper undergraduate and graduate levels. More advanced theoretical topics and greater depth of topics could be used in graduate courses. If all of the chapters in the book are covered in depth the book could be used in a two semester (i.e. eight month) course in environmetrics. However, as explained in the next section, there are various routes that can be followed for studying a useful subset of the chapters. Hence, the book could be used in a variety of specially designed one semester environmetrics courses. Exercises are presented at the end of all of the chapters.

Virtually all of the tools presented in the book are highly developed from a theoretical viewpoint and possess appropriate algorithms that permit them to be applied in practice. Therefore, the **methods are completely operational** and can be used now for solving actual problems in environmetrics. **Practitioners** who are studying specific environmental problems can refer to appropriate environmetrics techniques given in the book that are immediately useful to them. Moreover, the **McLeod-Hipel Time Series Package** (McLeod and Hipel, 1992) described in Section 1.7 is a computerized **decision support system** that can be employed by practitioners for applying appropriate environmetrics technologies to their problems and obtaining useful information upon which optimal decisions can be made.

The book holds a treasure house of ideas for **researchers** in environmetrics and time series modelling. More specifically, besides defining useful statistical tools as well as informative applications of the methods to actual data in order to explain clearly how to use them, the book puts the relative importance of environmetrics techniques into proper perspective. Furthermore, based upon both practical and theoretical needs, the book provides guidance as to where further worthwhile research is required.

In the next subsection, the overall **layout** of the book is described and **possible routes** for exploring the countryside of ideas contained in the chapters are traced out. Subsequently, in Section 1.6.3 the book is compared to other available books that deal with specific areas of environmetrics.

1.6.2 A Traveller's Guide

There are many different kinds of environmetrics techniques which can be applied to a wide variety of environmental problems. Consequently, one could envision a substantial number of sequences in which to present topics in environmetrics. For example, one could subdivide topics in environmetrics according to types of application problems. Another approach is to present the statistical methods from simpler to more complex and then to present applications later. The motivation for the **sequence of presentation of topics in this book is pedagogical.** First, it is assumed that a reader of this book has the background acquired after completing one introductory course in probability and statistics. Next, the order of topics in the book is for a reader who is studying environmetrics for the first time. Accordingly, the topics in time series modelling are arranged from simpler to more complex. Throughout the book, practical applications are given so the reader can appreciate how the methods work in practice and the insights that can be gained by utilizing them. After the reader has accumulated a variety of environmetrics tools in the earlier chapters, later in the book more complex environmental impact assessment studies are examined and general methodologies are described for systematically applying appropriate statistical methods from the toolbox of ideas that are available. Indeed, in Part X, general approaches are presented for trend assessment of ''messy'' environmental data.

Of course, many readers of this book may already have some background in environmetrics and, more specifically, time series modelling. These readers may wish to skip some of the earlier topics and immediately start with subjects presented in later sections in the book. Because readers may have varying backgrounds and reasons for studying environmetrics, there are **many different routes** through which one could tour the territory of environmetrics topics given in this book. The purposes of this section are to outline the topics covered in the book and suggest some good itineraries for the environmetrics traveller, depending upon what types of interesting tourist spots he or she would like to discover.

As can be seen from the Table of Contents, the **24 Chapters** in the book are divided into **10 main Parts**. The titles of the Chapters within each Part provide guidance about the topics given in each Part. For the convenience of the reader, Table 1.6.1 furnishes a tabulation of the titles of each of the ten Parts, along with a listing of the Chapter titles. Readers may also wish to refer to the brief one or two page descriptions given at the start of each Part where it first appears in the book. Finally, an overview of the book is also provided in the Preface.

Within **Part I**, labelled Scope and Background Material, Chapter 1 puts the overall objectives of the book into proper perspective and points out its main contributions to environmetrics. In particular, as explained in Section 1.2, statistics provides a powerful means for enhancing the **scientific method** in the search for solutions to pressing environmental problems. This book emphasizes the use of time series and other statistical methods for carrying out systematic data analysis studies of environmental time series. As pointed out in Section 1.2.4 and described in detail in Part X, a data analysis study consists of the two main stages of **exploratory data analysis** plus **confirmatory data analysis**. When fitting a specific time series model to a sequence of observations at the confirmatory data analysis stage, one can follow the identification, estimation and diagnostic check stages of **model construction** (Section 1.3). Moreover, in a given environmetrics study, one should keep in mind the overall **physical aspects** of the environmental problem (Section 1.4) as well as the influence of scientific studies upon the **decision making process** (Section 1.5). In the second chapter of Part I, some **basic statistical concepts** that are particularly useful in time series modelling are presented.

In order to provide the reader with specific tools that can be used in a data analysis study, **Part II** of the book describes some simple, yet well designed, time series models for fitting to yearly time series. Chapter 3 defines **AR** (autoregressive), **MA** (moving average) and **ARMA** (autoregressive-moving average) models for fitting to **stationary nonseasonal time series**. As explained in Section 2.4, stationarity means that the basic statistical properties of the series do not change over time. In Chapter 4, the **ARIMA** (autoregressive integrated moving average) family of models is defined for describing **nonstationary yearly time series**, where, for example, the mean levels may increase with time. As is the case with all of the models presented in the book, each model in Part II is clearly defined, its main theoretical properties that are useful in practical applications are pointed out, and illustrative examples are employed to explain how to fit the model to real data.

Table 1.6.2 gives a list of all of the time series models presented in the book, acronyms used to describe the models, locations in the book where model definitions, model construction and applications of the models can be found, as well as brief descriptions of the domains of applicability of the models. **The nonseasonal time series models of Part II actually form the theoretical foundations upon which more complicated models of later Parts can be defined.** However, before defining more models after Part II, the manner in which the models of Part II

Table 1.6.1. Parts and chapters in the book.

Part Numbers	Part Titles	Chapter Numbers	Chapter Titles
I	Scope and Background Material	1. 2.	Environmetrics, Science and Decision Making Basic Statistical Concepts
II	Linear Nonseasonal Models	3. 4.	Stationary Nonseasonal Models Nonstationary Nonseasonal Models
III	Model Construction	5. 6. 7.	Model Identification Parameter Estimation Diagnostic Checking
IV	Forecasting and Simulation	8. 9.	Forecasting with Nonseasonal Models Simulating with Nonseasonal Models
V	Long Memory Modelling	10. 11.	The Hurst Phenomenon and Fractional Gaussian Noise Fractional Autoregressive-Moving Average Models
VI	Seasonal Models	12. 13. 14. 15.	Seasonal Autoregressive Integrated Moving Average Models Deseasonalized Models Periodic Models Forecasting with Seasonal Models
VII	Multiple Input-Single Output Models	16. 17. 18.	Causality Constructing Transfer Function-Noise Models Forecasting with Transfer Function-Noise Models
VIII	Intervention Analysis	19.	Building Intervention Models
IX	Multiple Input-Multiple Output Models	20. 21.	General Multivariate Autoregressive-Moving Average Models Contemporaneous Autoregressive-Moving Average Models
X	Handling Messy Environmental Data	22. 23. 24.	Exploratory Data Analysis and Intervention Modelling in Confirmatory Data Analysis Nonparametric Tests for Trend Detection Regression Analysis and Trend Assessment

are systematically fitted to data is explained in **Part III**. Specifically, the identification, estimation and diagnostic check stages of **model construction** are clearly explained using practical applications. Later in the book, **these basic model building approaches of Part III are extended and modified for use with all the other time series models presented in Parts V to IX.**

Two major types of application of time series models are forecasting and simulation. In **Part IV**, procedures are presented for forecasting (Chapter 8) and simulating (Chapter 9) with the linear nonseasonal models of Part II. The objective of **forecasting** is to obtain the best possible estimates or forecasts of what will happen in the future based upon the time series model fitted to the historical time series as well as the most recent observations. When operating a system of reservoirs for producing hydroelectrical power, forecasts of the inflows to the reservoirs are used for developing an optimal operating policy which maximizes profits from the sale of the electricity. In **simulation**, time series models are utilized to produce possible future sequences of the phenomenon being modelled. Simulation can be used for designing a large scale engineering project and for studying the theoretical properties of a given time series model.

All of the time series models presented in this book can be used for forecasting and simulation. As a matter of fact, extensive experiments in forecasting and simulation given in many parts of the book demonstrate the great import of forecasting and simulation in environmetrics as well as the ability of ARMA-type models to perform better than their competitors. Table 1.6.3 lists the locations in the book where the forecasting and simulation procedures and applications are given for many of the models given in Table 1.6.2.

Long memory models were developed within the field of stochastic hydrology in an attempt to explain what is called the **Hurst Phenomenon**. Within Chapters 10 and 11 of **Part V**, two long memory models called **FGN** (Fractional Gaussian noise) and **FARMA** (fractional autoregressive-moving average) models, respectively, are defined for fitting to annual geophysical time series. Additionally, the Hurst phenomenon is defined in Chapter 10 and a proper explanation for the Hurst phenomenon is put forward. More specifically, it is demonstrated using extensive simulation experiments in Chapter 10, that **properly fitted ARMA models can statistically preserve historical statistics related to the Hurst phenomenon.**

By the end of Parts IV or V, the reader has a solid background in nonseasonal time series modelling. He or she knows the basic definitions of some useful yearly models (Part II and also Part V), understands how to apply these models to actual data sets (Part III), and knows how to calculate forecasts and simulated sequences with these models (Part IV). The reader is now in a position to appreciate how some of these nonseasonal models can be extended for use with other kinds of data. As indicated in Tables 1.6.1 and 1.6.2, models for fitting to seasonal data are described in Part VI immediately after the Part on long memory models. The three types of **seasonal models** described in Chapters 12 to 14 are the **SARIMA** (seasonal autoregressive integrated moving average), deseasonalized, and **PARMA** (periodic autoregressive-moving average) models, respectively. A special case of the PARMA models of Chapter 14 is the set of **PAR** (periodic autoregressive) models. In order to reduce the number of parameters in a PAR model, the **PPAR** (parsimonious periodic autoregressive model) can be employed. The deseasonalized and periodic models are designed for fitting to natural time series in which statistical characteristics such as the mean and variance must be accounted for in each season. Furthermore, periodic models can describe autocorrelation structures which change across the seasons. **Forecasting experiments** in Chapter 15 clearly demonstrate that PAR models are well suited for forecasting seasonal riverflows. SARIMA models can be used for forecasting nonstationary socio-economic time series such as water demand and electricity consumption.

In many natural systems, a single output or response variable is driven by one or more input or covariate series. The **TFN** (transfer function-noise) model of Part VII is designed for stochastically modelling the dynamic relationship between the input series and the single

Table 1.6.2. Time series models presented in the book.

Categories	Model Names	Acronyms	Definitions	Model Construction	Applications	Domains of Applicability
Linear Nonseasonal Models (Parts II and III)	Autoregressive	AR	Section 3.2	Part III	Introduced in Part II and completed in Part III.	Stationary annual time series.
	Moving average	MA	Section 3.3			
	Autoregressive - moving average	ARMA	Section 3.4			
	Autoregressive integrated moving average	ARIMA	Section 4.3		Section 4.3.3 Part III	Nonstationary yearly time series.
Long Memory Models (Part V)	Fractional Gaussian noise	FGN	Section 10.4.2	Sections 10.4.3 & 10.4.4	Section 10.4.7	Stationary annual series having long memory.
	Fractional autoregressive-moving average	FARMA	Section 11.2.2	Section 11.3	Section 11.5	
Seasonal Models (Part VI)	Seasonal autoregressive integrated moving average	SARIMA	Section 12.2	Section 12.3	Sections 12.4 and 14.6	Seasonal time series having non-stationarity within each season.
	Deseasonalized	DES	Section 13.2	Section 13.3	Sections 13.4 and 14.6	Seasonal time series for which mean and/or variance are preserved in each season.
	Periodic autoregressive	PAR	Section 14.2.2	Section 14.3	Section 14.6	Seasonal time series for which correlation structure varies across the seasons.
	Periodic autoregressive-moving average	PARMA	Section 14.2.3	Section 14.7		
	Parsimonious periodic autoregressive	PPAR	Section 14.5.2	Section 14.5.3	Section 14.6	

Cont'd...

Table 1.6.2. Time series models presented in the book (continued).

Categories	Model Names	Acronyms	Definitions	Model Construction	Applications	Domains of Applicability
Single Output and Multiple Input Models (Part VII)	Transfer function - noise	TFN	Sections 17.2 (single input) and 17.5.2 (multiple inputs)	Sections 17.3 (single input) and 17.5.3 (multiple inputs)	Sections 17.4 (single input) and 17.5.4 (multiple inputs)	Nonseasonal time series and seasonal data which are usually first deseasonalized.
Single Output Having Multiple Interventions, Missing Data, and Multiple Input Series (Part VIII)	Intervention		Section 19.2.2 (multiple interventions)	Section 19.2.3 (multiple interventions)	Sections 19.2.4 and 19.2.5 (multiple interventions)	Nonseasonal data and seasonal time series which are usually first deseasonalized. Mean level of output series can be affected by one or more intervention series and missing values in the output series can be estimated. Can also handle multiple input series.
			Section 19.3.3 (missing data)	Section 19.3.4 (missing data)	Sections 19.3.5 and 19.3.6 (missing data)	
			Section 19.4.2 (multiple interventions and missing data)	Section 19.4.3 (multiple interventions and missing data)	Sections 19.4.4 and 19.4.5 (multiple interventions and missing data)	
			Section 19.5.2 (multiple interventions, missing data and input series)	Section 19.5.3 (multiple interventions, missing data and input series)	Section 19.5.4 (multiple interventions, missing data and input series)	
Multiple Ouput and Multiple Input Models (Part IX)	General multivariate autoregressive-moving average	Multivariate ARMA	Section 20.2	Section 20.3		Nonseasonal and and deseasonalized time series for which there is feedback between series.
	Contemporaneous autoregressive-moving average	CARMA	Section 21.1	Section 21.3 and Appendix A21.1	Section 21.5	

Table 1.6.3. Forecasting and Simulation.

Model Names	Forecasting		Simulation	
	Algorithms	Applications	Algorithms	Applications
AR	Minimum mean square error forecasts defined in Section 8.2	Section 8.3	Sections 9.2, 9.3, 9.4, 9.6, and 9.7	Sections 9.8, 10.5 and 10.6
MA				
ARMA				
ARIMA			Sections 9.2 to 9.7	Section 9.8
FGN	Section 10.4.5	Section 8.3.4	Section 10.4.6	Section 10.5.3
FARMA	Section 11.4.3	Section 8.3.4	Section 11.4.2	
SARIMA	Section 15.2.2	Sections 15.3 and 15.4	Section 9.5	
DES	Sections 13.5 and 15.2.3		Section 13.5	
PAR	Sections 14.8 and 15.2.4		Section 14.8	Section 14.8.2
PARMA				
PPAR		Section 15.4		Section 14.8.2
TFN	Section 18.2	Sections 18.3 and 18.4	Section 18.5.3	
CARMA			Section 21.4 and Appendix A 21.2	

response. In **stochastic hydrology**, one may wish to model formally the manner in which precipitation and temperature cause riverflows. Qualitatively, a TFN model for describing this dynamic situation is written as:

Riverflows = Precipitation + Temperature + Noise

where the noise term is modelled as an ARMA model from Chapter 3. An example of the design of a TFN model for use in **stochastic water quality modelling** is:

Water	**=**	**Riverflows**	**+**	**Other**	**+**	**Noise**
Quality				**Water**		
Variable				**Quality**		
(ex. Phosphorous)				**Variables**		
				(ex. Water		
				Temperatures		
				and Turbidity)		

Because the above qualitative equation contains both water quantity and water quality time series, **the TFN model, as well as the related intervention model of Part VIII, provide a formal means for connecting stochastic hydrology (i.e., water quantity modelling) with statistical water quality modelling.**

As shown in the Table of Contents and Table 1.6.1, there are three chapters in Part VII. Chapter 16 explains how the **residual cross-correlation function** can be used to detect different kinds of **causality** between two variables. In Chapter 17, the **TFN model** is formally defined and flexible model building procedures along with illustrative applications are presented. The **forecasting experiments** of Chapter 18 clearly demonstrate that TFN models provide more accurate forecasts than their competitors. In fact, one of the forecasting experiments shows that **a simple TFN model forecasts better than a very complicated and expensive conceptual model.**

The **intervention model of Part VIII** constitutes a special type of TFN that is especially well suited for use in **environmental impact assessment**. Qualitatively, the intervention model has the following form:

Output	=	**Multiple**	+	**Multiple**	+	**Missing**	+	**Noise**
Variable		**Inputs**		**Interventions**		**Data**		

In addition to describing the effects of multiple input series upon a single response variable, the intervention model can simultaneously model the effects of one or more external interventions upon the mean level of the output series, estimate missing observations and handle correlated noise through an ARMA noise component. For the case of a **water quality application,** an intervention model could be written as:

Water Quality Variable = Riverflows + Other Water Quality Variables + Multiple Intervenions + Missing Data + Noise

The formal modelling and analysis of a data set using the intervention model is popularly referred to as **intervention analysis**. As pointed out in Section 1.2.4 and explained in detail in Section 19.4.5, a special form of the above intervention model can be designed for formally describing the drop in the mean level of the phosphorous series shown in Figure 1.1.1. Applications of a variety of intervention models to stochastic hydrology and environmental engineering data sets in Chapter 19 and Section 22.4, demonstrate that **the intervention model is one of the most comprehensive and flexible models available for use in environmetrics.** A sound and flexible theoretical design coupled with comprehensive model construction methods permit the intervention model to be conveniently and expeditiously applied in practice.

When there is **feedback** in a system, the input affects the output but the output can in turn have a bearing upon the input. For example, consider the situation where rivers drain into a large lake. The flow in a given river is caused by precipitation. However, evaporation from the large lake which is filled by the rivers causes precipitation that once again creates riverflows. To model formally this type of situation, one can employ the **multivariate ARMA** family of models of **Part IX** that has the form:

Multiple Outputs = Multiple Inputs + Noise

The terminology multivariate is employed because there are multiple output series. A drawback of the multivariate ARMA models is that they contain a great number of parameters. To reduce significantly the number of model parameters, one can employ the **CARMA** (contemporaneous ARMA) multivariate model of Chapter 21 for special types of applications. Good model construction methods, including a parameter estimation method that is efficient both statistically and computationally, are available for use with CARMA models.

A major strength of this book is the presentation of a variety of useful techniques that can be employed in complex environmental impact assessment studies. The objective of **Part X** is to explain how a variety of statistical methods can be used for **detecting and modelling trends**, as well as other statistical properties, in both water quantity and water quality time series. This is carried out within the overall framework of exploratory and confirmatory data analyses referred to in Sections 1.2.4 and 22.1. Within Section 22.3 a variety of informative **graphical tools** are described for use as **exploratory data analysis** tools. Additionally, in Section 24.2.2 it is explained how a technique called **robust locally weight regression smoothing** can be employed for tracing a trend that may be present in a time series. At the **confirmatory data analysis** stage, the three statistical approaches that are employed consist of **intervention analysis** (Chapter 19 and Section 22.4), **nonparametric tests** (Chapter 23) and **regression analysis** (Section 24.2.3).

Within a given data analysis study, one must select the most appropriate exploratory and confirmatory data analysis tools in order to discover, model and analyze the important statistical properties of the data. In fact, **data analysis** is composed of both an art and a craft. The **craft** consists of a knowledge and understanding of the main types of statistical tools that are available. This book, for example, describes and explains the capabilities of a wide variety of statistical methods. The **art** of data analysis is using the most appropriate statistical methods in an innovative and efficient manner for solving the data analysis problems currently being addressed. **The best way to explain how the art and craft of data analysis are carried out in practice is through the use of comprehensive real world case studies.**

Three major data analysis studies are presented in **Part X** of the book for carrying out trend assessments of water quality and water quantity time series. Each of the studies requires the development of a methodological approach within the encompassing structure of exploratory and confirmatory data analyses. Table 1.6.4 provides a list of the **three trend analysis methodologies** presented in Part X, the types of data analysis problems each procedure is applied to, brief descriptions of the methodologies, and the section numbers where explanations are provided. Notice that the first and third approaches in Table 1.6.4 deal with trend assessment of water quality and water quantity time series measured in rivers while the second procedure is concerned with water quality observations taken from a large lake. In all three studies, informative graphical techniques are employed as **exploratory data analysis** tools. At the **confirmatory data analysis** stage, different statistical methods are employed for trend modelling. Consider the first trend assessment methodology listed in Table 1.6.4. After filling in missing observations using the approach of Section 22.2, the time series approach of **intervention analysis** (Part VIII) is used to model trends in water quality series trends due to cutting down a forest in the Cabin Creek river basin of Alberta, Canada. As a matter of fact, the general form of one of the intervention models developed to describe the problem studied in Section 22.4.2 is:

Cabin	=	**Monthly**	+	**Cabin**	+	**Middle**	+	**Noise**
Creek		**Interventions**		**Creek**		**Fork**		
Water				**Flows**		**Water**		
Quality						**Quality**		
Series						**Series**		

Because the trees were not cut down in the nearby Middle Fork river basin, the water quality series from this river can account for changes in the same Cabin Creek water quality series that are not due to cutting down the trees. The estimate for the intervention parameter for a given month in the intervention component in this general intervention model provides an estimate for the change in the mean level of the water quality variable for the Cabin Creek for the month. Moreover, this model also stochastically accounts for the influence of riverflows upon the water quality variable in the Cabin Creek.

The environmental data used in the second and third studies in Table 1.6.4 are very "**messy**" because they possess undesirable properties such as having many missing values and possessing a large number of outliers. Consequently, **nonparametric methods** are used as confirmatory data analysis tools in these two studies for trend detection as well as other purposes. As explained in Chapter 23, nonparametric techniques have fewer underlying assumptions than competing parametric approaches and, therefore, are better designed for use with messy data. In fact, because nonparametric methods can be used for investigating a wide range of statistical characteristics that may be embedded in messy water quality data, Chapter 23 as well as some parts of Section 24.3 are dedicated to explaining how nonparametric tests can be effectively employed in environmental impact assessment studies. The main emphasis in the discussions of nonparametric methods is their use in **trend detection and modelling.** A review of **hypothesis testing** using nonparametric or parametric test statistics is presented in Section 23.2.

The third approach in Table 1.6.4 employs the regression analysis method called **robust locally weight regression smooth** for clearly tracing trends in time series plots of messy water quality data. Besides being used in graphical procedures at the exploratory data analysis stage, **regression analysis** can also be employed as a confirmatory data analysis method for a wide variety of purposes, including the estimation of the shapes and magnitudes of trends. For example, the last entry in Table 23.5.1 summarizes how regression analysis is employed in the Lake Erie water quality study of Chapter 23. Even though regression analysis usually constitutes a parametric technique, it can be used with data that are unevenly spaced.

The intervention model as well as the other types of time series analysis models presented in this book assume that observations are available at evenly spaced time intervals. If there are gaps in the data, then one must estimate the missing observations before fitting a time series model to the data set. Table 1.6.5 lists the **data filling techniques** described in this book as well as the main types of situations in which they can be used. A general discussion of estimating missing values is presented in Section 19.3 along with a detailed description of the intervention analysis approach to data filling.

Table 1.6.1 provides a summary of the ten main topics or Parts into which the book is divided. Each Part, in turn, is subdivided into a number of chapters. The **contents of each chapter** consist of the **introduction, main sections, conclusions, appendices, problems, and references.** Within the main sections of each chapter, any statistical methods that are described are usually accompanied by **practical environmental applications** so the reader can appreciate their usefulness in environmetrics.

Notice in Table 1.6.1 that the **model construction methods** for the linear time series models of Part II are presented in Part III. However, for each of the other time series models of Table 1.6.2, the model building procedures are usually given just after the model is defined. **Illustrative applications** are always used for demonstrating how a given time series model is fitted to a data set using appropriate model building techniques. The **types of environmental time series** used in the applications include water quantity, water quality, precipitation, ambient temperature, as well as miscellaneous series such as tree ring indices and mud varve thicknesses.

Depending upon the background of the reader, there are a **variety of different routes** he or she can follow when travelling through the environmetrics terrain given in the book. A neophyte in environmetrics may wish to follow sequentially the entire itinerary summarized in Table 1.6.1 in a **two semester course.** A **one semester course** could cover Part I, Chapter 3, Part III, Chapter 8, Chapters 13 and 14 in Part VI, Chapters 16 and 17 in Part VII, plus some Parts of Chapter 19. Someone who already has a background in basic ARMA modelling may wish to extend his knowledge by studying the more complex types of time series models given in Parts VII to IX and listed in Table 1.6.2. A course on **statistical environmental impact assessment** should concentrate on Parts VII, VIII and X. Someone who is mainly interested in **stochastic hydrology** should not miss studying Part V. Courses that emphasize **forecasting and simulation** can cover the topics listed in Tables 1.6.3.

Table 1.6.4. Trend assessment methodologies.

Methodologies	Kinds of Data	Descriptions	Sections
Trend Assessment Using Intervention Analysis (Chapter 22)	Water quality and water quantity time series measured in rivers.	Intervention analysis is used to describe the effects of cutting down a forest upon the mean levels of water quality and quantity time series. A seasonal adjustment data filling method is used to estimate missing values prior to fitting the intervention models.	Section 22.2 describes the seasonal adjustment data filling method. The exploratory data analysis is done in Section 22.3 while the intervention modelling is carried out in Section 22.4.
Trend Analysis of Water Quality Data Measured in Lakes (Chapter 23)	Water quality observations from a large lake.	Nonparametric trend tests and other statistical methods are used to detect trends in water quality variables in a lake that may be affected by nearby industrial developments.	Exploratory and confirmatory data analysis results are presented in Section 23.5. Table 23.5.1 lists all of the statistical methods used in the study.
Trend Analysis of Messy Water Quality Time Series Measured in Rivers (Chapter 24)	Messy water quality data and water quantity time series measured in rivers.	Procedures are presented for accounting for the effects of flow upon a given water quality series and eliminating any trend in the flow before its effect upon the water quality series is removed. The Spearman partial rank correlation test is used to detect trends in water quality time series when the effects of seasonality are partialled out.	Section 24.2.2 describes the robust locally weighted regression smooth for tracing trends. In Section 24.3, the overall trend analysis methodology is presented and applied.

Table 1.6.5. Data filling methods described in the book.

Techniques	Purposes	Locations
Intervention Analysis	The intervention model can be used to estimate missing observations simultaneously with other parameters in the intervention model. No more than about 5% of the observations in a time series should be missing when using the intervention model for data filling.	Chapter 19 is dedicated to describing and applying various versions of the intervention model. Section 19.3 explains in detail how the intervention model is used for filling in data. Besides Chapter 19, other applications of intervention analysis are presented in Section 22.4.2.
Seasonal Adjustment	The seasonal adjustment algorithm is designed for estimating missing values in seasonal time series when there is a great number of missing observations and there may be long time periods over which no measurements were taken.	The seasonal adjustment algorithm is defined and applied in Section 22.2.
Back Forecasting	A TFN model is used to connect two or more time series where the series overlap in time. Then the calibrated TFN model is used to "back forecast" the unknown values of the shorter series given as the response variable.	The TFN model along with model building techniques are presented in Chapter 17, while the technique of back forecasting is described in Section 18.4.

1.6.3 Comparisons to Other Available Literature

As mentioned in Section 1.1, this is a book about **environmetrics**. The techniques and methodological approaches presented in this book draw upon developments in the fields of **statistics**, **stochastic hydrology** and **statistical water quality modelling**. One purpose of this section is to point out some of the main books, journals and other literature from these three areas that can provide valuable **complementary reference material** for the reader. Within each chapter, **comprehensive references** are provided for the specific topics contained in the chapter. A second objective of this section is to **compare** the main contents of this book to other available literature. As is clearly explained, **this book constitutes a unique contribution to environmetrics.**

From the **field of statistics**, the major type of models used in this book is a wide variety of **time series analysis models.** Because time series analysis is employed extensively in many fields, a vast body of literature has evolved, especially during the past three decades when the advent of the electronic computer made it possible for both simple and complex time series models to be conveniently applied to large data sets. The **seminal textbook publication** which furnishes a systematic and comprehensive presentation of many time series models is the book of Box and Jenkins called "Time Series Analysis: Forecasting and Control." The first edition of their book appeared in 1970 while the second one was published in 1976. Besides defining useful time series models such as ARMA, ARIMA, SARIMA and TFN models, Box and Jenkins explain how to fit time series models to data sets by following the identification, estimation and diagnostic check stages of model construction. As a result, time series modelling became widely accepted for applying to practical problems in many different fields. For example, time series analysis has been widely used in **economics** for forecasting economic time series (Nelson, 1973; Montgomery and Johnson, 1976; Granger and Newbold, 1977; Firth, 1977; Makridakis and Wheelwright, 1978; Granger, 1980; Abraham and Ledolter, 1983; Pankratz, 1983) and in **electrical engineering** for estimating the state of a system in the presence of additive noise (Ljung, 1987; Haykin, 1990). In fact, because the time series analysis work of Box and Jenkins (1976) has become so widely adopted, many books, including this one, employ the notation of Box and Jenkins when defining time series models. Besides the book of Box and Jenkins (1976), **textbooks by statisticians** such as Jenkins and Watts (1968), Hannan (1970), Anderson (1971), Kendall (1973), Brillinger (1975), Chatfield (1975), Fuller (1976), Jenkins (1979), Priestley (1981), Pandit and Wu (1983), McLeod (1983), Vandaele (1983), Young (1984), and Brockwell and Davis (1987) furnish in a pedagogical fashion well explained accounts of developments in time series analysis. Within the statistical literature, the major **journals** to refer to for research and review articles include the Journal of the Royal Statistical Society (Series A, B and C), Biometrika, Journal of the American Statistical Association, the Annals of Statistics, Journal of Time Series Analysis, International Journal of Forecasting, and Communications in Statistics. Additionally, **proceedings from conferences** hosted by statistical societies and also by individuals, provide other valuable sources for research material on time series analysis. McLeod (1987), for example, edited a book of conference papers on stochastic hydrology. Anderson has edited 12 conference proceedings since 1976 (see, for instance, Anderson (1979) and Anderson et al. (1985)).

A number of **books** on **time series analysis in hydrology and water resources** have been written by authors such as Fiering (1967), Yevjevich (1972), Clarke (1973), Kottegoda (1980), Salas et al. (1980), Bras and Rodriguez-Iturbe (1985), and McCuen and Snyder (1986). A monograph edited by Hipel (1985) contains many original contributions on time series analysis in water resources. Both the water resources systems book of Loucks et al. (1981) and the stochastic modelling text of Kashyap and Rao (1976) contain some chapters on the modelling of hydrological time series. Moreover, the book of Helsel and Hirsch (1992) on statistical methods in water resources has chapters on exploratory data analysis, regression analysis and trend tests. Although the measurement and subsequent analysis of water resources time series started a long time ago, the proliferation of time series analysis research in water resources commenced in the early 1960's. In fact, the **journals** Water Resources Bulletin and Water Resources Research, founded at that time by the American Water Resources Association and the American Geophysical Union, respectively, have published papers on time series analysis throughout their history. Other water resources journals that present applications and theoretical developments in time series analysis, include Stochastic Hydrology and Hydraulics, the Journal of Hydrology, Journal of the Water Resources Planning and Management Division of the American Society of Civil Engineers (ASCE), Journal of the Hydraulics Division of ASCE, Advances in Water Resources and Hydrological Sciences Bulletin. Moreover, proceedings from water resources conferences, such as the ones edited by Shen (1976), McBean et al. (1979a,b) and Shen et al. (1986), provide a rich variety of papers on time series analysis in water resources. In addition to time series models, the international conference on **Stochastic and Statistical Methods in Hydrology and Environmental Engineering** held at the University of Waterloo from June 21 to 23, 1993, had many paper presentations regarding the other kinds of stochastic models shown in Table 1.4.1. This conference was held in the honour and memory of the late Professor T. E. Unny.

In addition to time series analysis methods from statistics and stochastic hydrology, this book also deals with ideas from **statistical water quality modelling.** Some of the models described in the stochastic hydrology books referred to in the previous paragraph can be applied to water quality time series. However, the **intervention model** of Chapters 19 and 22 in this book as well as the **TFN model** of Part VII, constitute time series models that are especially well designed for simultaneously modelling both water quality and quantity series. When the environmental data are quite messy and there are a great number of missing values, one may have to employ **nonparametric tests** (Kendall, 1975) in an environmental impact assessment study. In his book on statistical methods for environmental pollution monitoring, Gilbert (1987) makes extensive use of nonparametric techniques. A monograph edited by Hipel (1988) on nonparametric approaches to environmental impact assessment contains papers having applications of nonparametric trend tests to water quality time series. Papers regarding the use of time series models and nonparametric tests in water quality modelling appear in many of the water resources journals referred to in the above paragraph. In Chapter 23 and Section 24.3 of this book a number of useful nonparametric trend tests are presented along with water quality applications. Other **environmental journals** having statistical papers on water quality modelling include Environmetrics, Journal of the Environmental Engineering Division of ASCE, Environmental Management, and Environmental Monitoring and Assessment. Additionally, a number of **conference proceedings** on statistical water quality modelling are available. For instance, El-Shaarawi and Esterby (1982), El-Shaarawi and Kwiatowski (1986) and Chapman and El-Shaarawi (1989) have edited conference proceedings on water quality monitoring and assessment.

Given the impressive array of literature available in time series analysis, stochastic hydrology and statistical water quality modelling, one is tempted to ask what are the **relative contributions** of this book. In fact, there are many ways in which this book combines, enhances and extends previous research. First, the current book **combines the aforementioned three areas in a coherent and systematic manner in order to have a comprehensive book on time series methods in environmetrics.** As a result, stochastic hydrology and statistical water quality modelling are no longer considered to be separate fields. Rather, techniques and approaches from both areas are used in combination with other statistical methods to confront challenging environmental problems. Secondly, this book contains the most **recent and useful developments in time series modelling.** As a matter of fact, virtually all of the time series and other statistical methods presented in the book are well developed theoretically and a range of flexible model construction algorithms are available to allow them to be immediately applied to real world data sets. Thirdly, the book contains **many theoretical contributions and applications** that have been developed by the authors and their colleagues. Although much of the research has appeared during the past fifteen years in various water resources and statistical journals, this is the first time that **it is published in a pedagogical fashion within a single document.** Fourthly, this book puts great emphasis on the use of both time series analysis and nonparametric methods in **environmental impact assessment** studies. Using challenging real world applications, Parts VII, VIII and X in the book clearly explain how this is accomplished in practice. Finally, as is explained in Section 1.6.2, the book is designed to be as flexible as possible so that it **can satisfy the specific needs of individual readers.** It could, for example, be used as a one semester course in environmetrics. A person who is solely interested in stochastic hydrology or statistical water quality modelling could refer to the appropriate chapters in the book, as pointed out in Section 1.6.2. The book could also be used as a main text on time series modelling within a statistics department.

1.7 DECISION SUPPORT SYSTEM FOR TIME SERIES MODELLING

To carry out a proper data analysis study, one requires the employment of a flexible **Decision Support System** (Sage, 1991). The **McLeod-Hipel Time Series (MHTS) Package** (McLeod and Hipel, 1992) constitutes a comprehensive decision support system for performing extensive data analysis investigations in order to obtain easily understandable results upon which sound decisions can be made.

As explained in Section 1.2.4, **data analysis** consists of both exploratory and confirmatory data analyses. A wide variety of **informative graphical methods** are contained in the MHTS package as **exploratory data analysis** tools to allow a user to clearly visualize the key statistical characteristics of the time series that he or she is studying. Some of the many graphical methods contained in the MHTS system are described in Section 22.3 as well as elsewhere in the book. Figure 1.1.1, for example, is a simple yet informative plot created by the MHTS package that clearly shows the step drop in the phosphorous levels in a river due to the introduction of tertiary treatment in upstream sewage plants.

At the **confirmatory data analysis** stage, the MHTS package can, for instance, fit a model to a data set in order to provide a more precise mathematical description of the main statistical properties of the data. For example, in Section 19.4, the MHTS package is employed for fitting the most appropriate **intervention model** to the time series displayed in Figure 1.1.1. Besides mathematically modelling the stochastic structure of the time series, the intervention model provides an efficient estimate of the magnitude of the drop in the mean level of the phosphorous series in Figure 1.1.1 after the intervention of introducing tertiary sewage treatment.

The MHTS package can be employed for fitting virtually all of the different kinds of time series models listed in Table 1.6.2 to time series. For each type of time series model, the three stages of **model construction** depicted in Figure 1.3.1 are adhered to when using the MHTS package. Moreover, the most appropriate and up-to-date tools are used by the MHTS package for each kind of model during the identification, estimation and diagnostic check stages of the model development.

The MHTS package allows a practitioner or researcher to employ calibrated time series models for carrying out **applications** such as **forecasting and simulation experiments.** Table 1.6.3, for example, summarizes a wide variety of forecasting and simulation investigations described in the book.

Each data analysis study usually contains unique problems and challenges that require specialized attention. The MHTS package furnishes the user with a **comprehensive array of tools** from which he or she can select the most appropriate ones in order to discover and build the most effective explanations and solutions to the problems being studied. This inherent comprehensive, yet flexible, design of the MHTS system is especially needed when dealing with the type of **messy environmental data** described in Part X of the book. Table 1.6.4, for instance, lists some general kinds of **trend assessment** studies that can be conviently executed in an iterative, yet logical and systematic fashion, using the MHTS package.

The MHTS package is extremely **user-friendly** and is designed for use by both novices and experts in time series modelling and analysis. Additionally, the MHTS package is **menu-driven** using attractive screen displays, operates on **personal computers** that use **DOS**, and can support a variety of dot-matrix and laser printers.

The MHTS package is probably the most advanced decision support system available which employs such a **rich range of ARMA-type, as well as other kinds of time series methods,** in statistical decision making. The MHTS package permits data analysis methods to be effectively utilized within the **scientific method** described in Section 1.2 and summarized in Figures 1.2.1 to 1.2.3, as well as the overall structure of **engineering decision making** explained in Section 1.5.1 and depicted in Figure 1.5.2. Finally, the MHTS package places a set of valuable statistical tools directly into the hands of decision makers working in **environmetrics who must make real decisions now about complex environmental problems.**

1.8 CONCLUDING REMARKS

The main purpose of this book is to present the art and craft of environmetrics for modelling water resources and environmental systems. The craft consists of a set of useful statistical tools while the art is composed of general procedures or methodologies for applying these tools to environmental time series. As pointed out in Section 1.6.2, the **tools of the trade** are stressed earlier in the book while **general methodologies** are presented after the reader has some tools to work with.

The **topics** on environmetrics given in this book and summarized in Table 1.6.1 are based upon **contributions from the fields of statistics, stochastic hydrology and statistical water quality modelling.** The main set of statistical methods used in the book is a range of families of **time series models.** Other statistical techniques include **graphical methods, nonparametric**

trend tests and **regression models.** Within the framework of the intervention model, for example, both water quality and water quantity data can be simultaneously considered in an environmental impact assessment study (see Chapters 19 and 22).

In Section 1.2, the **scientific method** is described and it is explained how statistics can enhance the scientific method so that good solutions to pressing environmental problems can be efficiently and expeditiously found. When carrying out a **scientific data analysis** study, informative exploratory and confirmatory data analysis tools can be employed. Graphical methods can be used as **exploratory data analysis** methods for discovering general statistical properties of a given data set. At the **confirmatory data analysis** stage, both parametric and nonparametric methods can be employed for rigorously modelling from a mathematical viewpoint the main statistical characteristics. For example, the intervention model from Part VIII can be utilized to model the shape and magnitude of the trend in the phosphorous data in Figure 1.1.1.

As listed in Table 1.6.2, a **wide variety of time series models** are presented in the book. In Section 1.3, it is explained how one should follow the identification, estimation and diagnostic check stages of **model construction** in order to design a parsimonious model that provides a reasonable statistical fit to the time series. In any statistical study, one should keep in mind the physical aspects of the environmental problem. Section 1.4 explains how the **hydrological cycle** forms a sound environmental system model for use in environmetrics.

In **environmental decision making**, one should remember how the results of an environmetrics study can influence the overall political decisions which are eventually made when a specific alternative solution to a given environmental problem is selected. As explained in Section 1.5, rigorous scientific studies can result in decisions that are environmentally acceptable. Moreover, active participation by scientists and engineers in both tactical and strategic decision making will allow better solutions to be reached for solving pressing environmental problems. An analogy with the success of Japanese industry will help to explain why this should be so. A majority of people at all levels of management, including the board of directors of most major Japanese corporations, have technical training as engineers and scientists (Reich, 1983). In addition, over a time period of a few decades these corporate leaders worked their way from the bottom of a given company to the top. Because they understand both the technical and social problems at all levels in the company, they are properly educated to make sound tactical and strategic decisions. The worldwide large sales of high quality Japanese products attest to the success of having wise leadership. Likewise, when it comes to solving complex environmental problems, properly educated decision makers are sorely needed. **The authors of this book feel that a good scientific education coupled with many years of solving tough environmental problems will produce leaders who will be capable of making optimal tactical and strategic decisions.** This environmental decision making may be carried out within a framework similar to that described in Section 1.5. One should always recall that good science can create imaginative solutions to environmental problems which in turn can influence the preferences of the decision makers who must eventually make the strategic decisions. Finally, by employing the **McLeod-Hipel Time Series Package** described in Section 1.7, a decision maker can use this flexible **decision support system** to immediately take full advantage of the rich array of environmetrics methods presented in the book.

Fertile fields of ideas in environmetrics await the curious reader in the upcoming chapters of this book. Depending upon the background and interests of the individual traveller, a **variety of touring routes** are suggested in Section 1.6.2. A reader who would like a review of some basic statistical concepts used in time series modelling may wish to start his or her journey by moving on to Chapter 2. Bon voyage!

PROBLEMS

1.1 What roles do you think science and statistics should play in developing sound environmental policies for restoring and preserving the natural environment?

1.2 Some serious types of environmental problems are referred to in Section 1.2.1. Discuss an important environmental problem that is of great concern to you. Outline realistic steps that could be taken to overcome this problem. Kindly provide the references from which you obtained your background information.

1.3 Many governments are deeply concerned with having human rights for each individual citizen. How do you think human rights and protection of the natural environment should be related? Discuss the situation in your own country.

1.4 The scientific method is explained in Section 1.2.2. Make a list of some of the key historical breakthroughs in the development of the scientific method. Wherever possible, point out when statistics played a key role in the improvement of the scientific method. Kindly provide a list of the references from which you obtained your material.

1.5 Tukey (1977) suggests a wide variety of graphical techniques for use as exploratory data analysis tools. Make a list of some of the main graphical procedures put forward by Tukey and briefly describe their purposes.

1.6 Explain in your own words why you think human beings like to develop models of natural and social systems. What is your opinion of the modelling procedure outlined in Section 1.3?

1.7 Chatfield (1988) describes a general approach for addressing real-life statistical problems. Summarize the main aspects of his approach and comment upon the advantages as well as the drawbacks of his methodology.

1.8 In the hydrological cycle displayed in Figure 1.4.1, the throughput to the system is water. Explain how energy could be used as the throughput in the hydrological cycle. Which throughput do you feel is more informative and easier to understand?

1.9 The Garrison Diversion Unit conflict of Sections 1.5.1 and 1.5.3 constitutes an example of an international environmental dispute for which good scientific and economic studies were carried out. Describe another planned large-scale project which is causing controversies to arise because of possible detrimental environmental effects. Who are the key decision makers involved in this environmental conflict and what are the main courses of action that each decision maker can follow. Discuss any scientific and/or economic studies that have been completed as well as how the results of the studies may influence any of the decision makers and the possible resolutions to the dispute. How are statistical methods used in any of these studies? If proper scientific or economic studies are not currently being executed, suggest how they could be done. Be sure to reference newspapers, magazines, journals and books from which you obtain your background information.

1.10 In Section 1.6.3, a number of water resources journals are mentioned. Select any two of the water resources journals and go to your library to obtain the last two years of publications. For each journal, make a list of categories under which the individual papers could be classified. What percentage of papers in each journal deal with environmetrics? Discuss the main types of environmetrics papers that are published.

1.11 A number of books on stochastic hydrology are mentioned in Section 1.6.3. Select one of these books and make a list of the major types of time series models discussed in the book. Compare this list of models to the one given in Table 1.6.2.

1.12 A decision support system for time series modelling and analysis is described in Section 1.7. By referring to appropriate literature, explain the basic design, function and user-friendly features of a decision support system in general. You may wish to read articles published in journals such as Decision Support Systems and Information and Decision Technologies, appropriate textbooks like the one by Sage (1991) and scientific encyclopediae such as the one edited by Sage (1990).

1.13 The McLeod-Hipel Time Series Package mentioned in Section 1.7 contains a wide variety of representative time series. Pick out a hydrological series that is of interest to you and use the package to produce a graph over time of the series. Write down some of the statistical characteristics of the time series that you can visually detect in the graph. If you feel adventurous, use the package to fit an appropriate time series model to the data and to produce forecasts 10 steps into the future. List some of the features that you like about the McLeod-Hipel package.

1.14 Why are education, in general, and science and engineering, in particular, so highly respected in Japan? What influence has this reverence for education had upon the economy of Japan? Have the Japanese adopted sound environmental policies within their own country and how are these policies connected to economic policies?

1.15 After you finish reading Chapter 1, write down your main reasons or objectives for studying environmetrics. After you complete taking an envionmetrics course using this book or reading the entire book on your own, refer to this list to see if your goals have been met.

REFERENCES

CONFLICT ANALYSIS

Fang, L., Hipel, K. W. and Kilgour, D. M. (1988). The graph model approach to environmental conflict resolution. *Journal of Environmental Management*, 27:195-212.

Fang, L., Hipel, K. W. and Kilgour, D. M. (1993). *Interactive Decision Making: The Graph Model for Conflict Resolution*. Wiley, New York.

Fraser, N. M. and Hipel, K. W. (1984). *Conflict Analysis: Models and Resolutions*. North-Holland, New York.

Hipel, K. W. (1990). Decision technologies for conflict analysis. *Information and Decision Technologies*, 16(3):185-214.

Hipel, K. W. and Fraser, N. M. (1980). Metagame analysis of the Garrison conflict. *Water Resources Research*, 16(4):629-637.

Howard, N. (1971). *Paradoxes of Rationality, Theory of Metagames and Political Behavior.* M.I.T. Press, Cambridge, Massachusetts.

von Neumann, J. and Morgenstern, O. (1953). *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, New Jersey, third edition.

DATA COLLECTION

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Harmancioglu, N. B. and Alpaslan, N. (1992). Water quality monitoring network design: A problem of multi-objective decision making. *Water Resources Bulletin* 28(1):179-192.

Lerner, D. (Editor) (1986). *Monitoring to Detect Changes in Water Quality Series*. Proceedings of a symposium held during the 2nd Scientific Assembly of the International Association of Hydrological Sciences (IAHS) held in Budapest, Hungary, in July, 1986, IAHS Publication No. 157.

Lettenmaier, D. P., Hipel, K. W. and McLeod, A. I. (1978). Assessment of environmental impacts, part two: Data collection. *Environmental Management*, 2(6):537-554.

Loftis, J. C., McBride, G. B. and Ellis, J. C. (1991). Considerations of scale in water quality monitoring and data analysis. *Water Resources Bulletin* 27(2):255-264.

Moss, M. E. (1979). Some basic considerations in the design of hydrologic data networks. *Water Resources Research*, 15(6):1673-1676.

Ward, R. C. and Loftis, J. C. (1986). Establishing statistical design criteria for water quality monitoring systems: Review and synthesis. *Water Resources Bulletin* 22(5):759-767.

Ward, R. C., Loftis, J. C. and McBride, G. B. (1986). The data-rich but information-poor syndrome in water quality monitoring. *Environmental Management* 10:291-297.

Ward, R. C., Loftis, J. C. and McBride, G. B. (Editors) (1989). *Design of Water Quality Information Systems*. Proceedings of an International Symposium held at Fort Collins, Colorado from June 7 to 9, 1989, sponsored by Colorado State University and the U. S. Environmental Protection Agency, Information Series No. 61, Colorado Water Resources Research Institute.

Whitfield, P. H. (1988). Goals and data collection designs for water quality monitoring. *Water Resources Bulletin* 24(4):775-780.

ECONOMIC FORECASTING

Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*. John Wiley, New York.

Firth, M. (1977). *Forecasting Methods in Business and Management*. Edward Arnold Publishers, London.

Granger, C. W. J. (1980). *Forecasting in Business and Economics.* Academic Press, New York.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series.* Academic Press, New York.

Makridakis, S. and Wheelwright, S. (1978). *Interactive Forecasting: Univariate and Multivariate Methods.* Holden-Day, San Francisco.

Montgomery, D. C. and Johnson, L. A. (1976). *Forecasting and Time Series Analysis.* McGraw-Hill, New York.

Nelson, C. R. (1973). *Applied Time Series Analysis for Managerial Forecasting.* Holden-Day, San Francisco.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models.* John Wiley, New York.

HYDROLOGY AND ENVIRONMENTAL SYSTEMS

Bennett, R. J. and Chorley, R. J. (1978). *Environmental Systems: Philosophy, Analysis and Control.* Princeton University Press, New Jersey.

Eagleson, P. S. (1970). *Dynamic Hydrology.* McGraw-Hill, New York.

Eagleson, P. S. (1991). Hydrologic science: a distinct geoscience. *Reviews of Geophysics,* 29(2):237-248.

Falkenmark, M. (1990). Environmental management - what is the role of the hydrologist? *Paper presented at the 1990 Commemorative Symposium for 25 Years of IHD (International Hydrological Decade) and IHP (International Hydrological Program).* UNESCO (United Nations Education, Scientific and Cultural Organization), held March 15-17, 1990, Paris, France.

Linsley, Jr., R. K., Kohler, M. A. and Paulhus, J. L. H. (1982). *Hydrology for Engineers.* McGraw-Hill, New York.

McCuen, R. H. (1989). *Hydrologic Analysis and Design.* Prentice-Hall, Englewood Cliffs, New Jersey.

National Research Council (1991). *Opportunities in the Hydrologic Sciences.* National Academy Press, Washington, D.C.

White, I. D., Mottershead, D. N. and Harrison, S. J. (1984). *Environmental Systems.* George Allen & Unwin, London.

OPERATIONAL RESEARCH AND SYSTEMS SCIENCES

Blackett, P. M. S. (1962). *Studies of War.* Oliver and Boyd, Edinburgh.

Checkland, P. (1981). *Systems Thinking, Systems Practice.* John Wiley, Chichester, United Kingdom.

Goicoechea, A., Hansen, D. R. and Duckstein, L. (1982). *Multiobjective Decision Analysis with Engineering and Business Applications.* Wiley, New York.

Hillier, F. S. and Lieberman, G. J. (1990). *Introduction to Operations Research.* McGraw-Hill, New York, fifth edition.

Hipel, K. W. (Editor) (1992). *Multiple Objective Decision Making in Water Resources*. AWRA Monograph #18 published by the American Water Resources Association, 5410 Grosvenor Lane, Suite 220, Bethesda, Maryland 20814-2192, U.S.A.

Loucks, D. P., Stedinger, J. R. and Haith, D. A. (1981). *Water Resources Systems Planning and Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.

Maas, A., Hufschmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A. and Fair, M. G. (1962). *Design of Water-Resource Systems*. Harvard University Press, Cambridge, Massachusetts.

Miser, H. J. and Quade, E. S. (Editors) (1985). *Handbook of Systems Analysis: Overview of Uses, Procedures, Applications, and Practice*. North-Holland, New York.

Miser, H. J. and Quade, E. S. (Editors) (1988). *Handbook of Systems Analysis: Craft Issues and Procedural Choices*. North-Holland, New York.

Radford, K. J. (1989). *Individual and Small Group Decisions*. Springer-Verlag, New York.

Reich, R. B. (1983). *The Next American Frontier*. Time Books, New York.

Rosenhead, J. (Editor) (1989). *Analysis for a Problematic World*. Wiley, Chichester, United Kingdom.

Roy, B. (1985). *Méthodologie Multicritère d'Aide a la Décision*. Economica, Paris.

Sage, A. P. (Editor) (1990). *Concise Encyclopedia of Information Processing in Systems and Organizations*. Pergamon, Oxford.

Sage, A. P. (1991). *Decision Support Systems Engineering*. Wiley, New York.

Vincke, P. (1992). *Multicriteria Decision-aid*. Wiley, Chichester, United Kingdom.

Waddington, C. H. (1973). *OR in World War 2*. Elek Science, London.

Wagner, H. M. (1975). *Principles of Operations Research*. Prentice-Hall, Englewood Cliffs, New Jersey, second edition.

SCIENCE

Box, J. (1978). *R.A. Fisher: The Life of a Scientist*. Wiley, New York.

Dransfield, J., Flenley, J. R., King, S. M., Harkness, D. D. and Rapu, S. (1984). A recently extinct palm from Easter Island. *Nature*, 312(5996):750-752.

Flenley, J. R. and King, S. M. (1984). Late quaternary pollen records from Easter Island. *Nature*, 307(5946):47-50.

Groen, J., Smit, E. and Eijsvoogel, J. (Editors) (1990). *The Discipline of Curiosity*. Elsevier, Amsterdam.

Levine, J. S. (1990). Global biomass burning: atmospheric, climatic and biospheric implications. *EOS, Transactions of the American Geophysical Union*, 71(37):1075-1077.

Russell, B. (1946). *History of Western Philosophy*. Allen and Unwin, London.

STATISTICAL WATER QUALITY MODELLING

Chapman, D. T. and El-Shaarawi, A. H. (Editors) (1989). Statistical methods for the assessment of point source pollution. In *Proceedings of a Workshop on Statistical Methods for the Assessment of Point Source Pollution,* held Sept. 12-14, 1988, at the Canada Centre for Inland Waters, Burlington, Ontario, Canada, Kluwer, Dordrecht, the Netherlands.

El-Shaarawi, A. H. and Esterby, S. R. (Editors) (1982). *Time Series Methods in Hydrosciences.* North-Holland, Amsterdam, The Netherlands.

El-Shaarawi, A. H. and Kwiatowski, R. E. (Editors) (1986). *Statistical Aspects of Water Quality Monitoring.* Elsevier, Amsterdam.

Hipel, K. W. (Editor) (1988). *Nonparametric Approaches to Environmental Impact Assessment.* American Water Resources Association, Bethesda, Maryland.

STATISTICS

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* AC-19:716-723.

Box, G. E. P. (1974). Statistics and the environment. *J. Wash. Acad. Sci.,* 64(2):52-59.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association,* 71(356):791-799.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 26:211-252.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters.* Wiley, New York.

Chatfield, C. (1988). *Problem Solving: A Statisticians Guide.* Chapman and Hall, London.

Helsel, D. R. and Hirsch, R. M. (1992). *Statistical Methods in Water Resources.* Elsevier, Amsterdam.

Hunter, J. S. (1990). Keynote address at the First International Conference on Statistical Methods for the Environmental Sciences, held April 4-7, Cairo, Egypt, 1989. *Environmetrics,* 1(1):1-6.

Kendall, M. G. (1975). *Rank Correlation Methods.* Charles Griffin, London, fourth edition.

McPherson, G. (1990). *Statistics in Scientific Investigations.* Springer-Verlag, New York.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts.

STOCHASTIC HYDROLOGY

Bras, R. L. and Rodriguez-Iturbe, I. (1985). *Random Functions and Hydrology.* Addison-Wesley, Reading, Massachusetts.

Clarke, R. T. (1973). *Mathematical Models in Hydrology.* Irrigation and Drainage Paper, Food and Agriculture Organization of the United Nations, Rome, 1973.

Fiering, M. B. (1967). *Streamflow Synthesis.* Harvard University Press, Cambridge, Massachusetts.

Hipel, K. W. (Editor) (1985). *Time Series Analysis in Water Resources*. American Water Resources Association, Bethesda, Maryland.

Kashyap, R. L. and Rao, A. R. (1976). *Dynamic Stochastic Models from Empirical Data*. Academic Press, New York.

Kottegoda, N. T. (1980). *Stochastic Water Resources Technology*. MacMillan Press, London.

McBean, E. A., Hipel, K. W. and Unny, T. E. (Editors) (1979a). *Inputs for Risk Analysis in Water Systems*. Water Resources Publications, Fort Collins, Colorado.

McBean, E. A., Hipel, K. W. and Unny, T. E. (Editors) (1979b). *Reliability in Water Resources Management*. Water Resources Publications, Fort Collins, Colorado.

McLeod, A. I. (Editor) (1987). Advances in statistical sciences. In *Festschrift in Honor of Professor V.M. Joshi's 70th Birthday, Volume IV, Stochastic Hydrology* (editors across all volumes are I.B. MacNeill and G.J. Umphrey), Reidel, Dordrecht, the Netherlands.

McCuen, R. H. and Snyder, W. M. (1986). *Hydrologic Modeling: Statistical Methods and Applications*. Prentice-Hall, Englewood Cliffs, New Jersey.

Pereira, M. V. F., Oliveira, G. C., Costa, C. C. G. and Kelman, J. (1984). Stochastic streamflow models for hydroelectric systems. *Water Resources Research*, 20(3):379-390.

Pereira, M. V. F. and Pinto, L. M. V. G. (1985). Stochastic optimization of a multireservoir hydroelectric system: A decomposition approach. *Water Resources Research*, 21(6):779-792.

Salas, J. D., Delleur, J. W., Yevjevich, V. and Lane, W. L. (1980). *Applied Modelling of Hydrologic Series*. Water Resources Publications, Littleton, Colorado.

Salas, J. D. and Smith, R. A. (1981). Physical basis of stochastic models of annual flows. *Water Resources Research*, 17(2):428-430.

Shen, H. W. (1976). *Stochastic Approaches to Water Resources, Volumes 1 and 2*. Colorado State University, Fort Collins, Colorado.

Shen, H. W., Obeysekera, J. T. B., Yevjevich, V. and Decoursey, D. G. (1986). Multivariate analysis of hydrological processes. In *Proceedings of the Fourth International Hydrology Symposium* held at Colorado State University, Fort Collins, Colorado, July 15-17, 1985. Engineering Research Center, Colorado State University.

Silva, L. F. C. A., Sales, P. R. H., Araripe Neto, T. A., Terry, L. A. and Pereira, M. F. V. (1984). *Coordinating the energy generation of the Brazilian system*. Technical report, Departamento de Operacao Energetica, Eletrobras, Rio de Janeiro, Brazil.

Yevjevich, V. (1972). *Stochastic Processes in Hydrology*. Water Resources Publications, Littleton, Colorado.

TIME SERIES ANALYSIS AND STOCHASTIC PROCESSES

Anderson, O. D. (Editor) (1979). *Proceedings of the International Time Series Meeting, Nottingham University, March 1979*. North-Holland, Amsterdam.

Anderson, O. D., Ord, J. K. and Robinson, E. A. (1985). Time series analysis: Theory and practice 6. In *Proceedings of the International Conference held at Toronto, Canada, August 10-14, 1983*. North-Holland, Amsterdam.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer-Verlag, New York.

Chatfield, C. (1975). *The Analysis of Time Series: Theory and Practice*. Chapman and Hall, London.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. John Wiley, New York.

Hannan, E. J. (1970). *Multiple Time Series*. John Wiley, New York.

Haykin, S. (1990). *Modern Filters*. MacMillan, New York.

Jenkins, G. M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. Gwilym Jenkins and Partners Ltd., Parkfield, Greaves Road, Lancaster, England.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Kendall, M. G. (1973). *Time Series*. Hafner Press, New York.

Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey.

McLeod, A. I. and Hipel, K. W. (1992). The McLeod-Hipel Time Series Package. Copyright of this decision support system owned by A.I. McLeod and K.W. Hipel. Distributed by McLeod-Hipel Research, 121 Longwood Drive, Waterloo, Ontario, Canada, N2L 4B6. Tel: (519) 884-2089.

McLeod, G. (1983). *Box-Jenkins in Practice, Volume 1, Univariate Stochastic and Transfer Function/Intervention Analysis*. Gwilym Jenkins and Partners Ltd., Parkfield, Greaves Road, Lancaster, England.

Pandit, S. M. and Wu, S. M. (1983). *Time Series and System Analysis with Applications*. John Wiley, New York.

Priestley, M. B. (1981). *Spectral Analysis and Time Series, Volume 1: Univariate Series, Volume 2: Multivariate Series, Prediction and Control*. Academic Press, London.

Vandaele, W. (1983). *Applied Time Series and Box-Jenkins Models*. Academic Press, New York.

Young, P. C. (1984). *Recursive Estimation and Time-Series Analysis*. Springer-Verlag, Berlin.

CHAPTER 2

BASIC STATISTICAL CONCEPTS

2.1 INTRODUCTION

In this chapter, the general properties of *time series* and *stochastic processes* are firstly discussed. This leads to the problem of deciding upon in which situations it is feasible to assume that the statistical characteristics of a time series under consideration are more or less constant over time and hence it is permissible to fit a stationary stochastic model to the data. A general appraisal is given regarding the controversies surrounding *stationarity* and *nonstationarity*. Following this, some *statistical definitions* are presented for examining stationary data in the *time domain* while the usefulness of the cumulative periodogram for *frequency domain* analyses is pointed out. Finally, the importance of linear stationary models in the environmental sciences is demonstrated by explaining the relevant results from the *Wold decomposition* theorem (Wold, 1954).

2.2 TIME SERIES

A *time series* is a set of observations that are arranged chronologically. In time series analysis, the order of occurrence of the observations is crucial. When a meteorologist wishes to forecast the weather conditions for tomorrow, the time sequence in which previous weather conditions evolved is of utmost importance. If the chronological ordering of the data were ignored, much of the information contained in the time series would be lost and the meteorologist would have a difficult task when attempting to forecast future weather patterns.

Data can be collected continuously over time. For example, temperature readings and the depth of a river may be recorded on a continuous graph. Data that are measured at every moment of time, constitute a *continuous time series*. Other types of observations may be recorded at discrete points in time and the resulting time series is said to be *discrete*. In certain situations, the time interval between sequential observations may vary. When the pollution levels in a river are being monitored downstream from a sewage treatment plant, readings may be taken every half hour during the daytime and once every two hours during the night when the pollutant concentrations fluctuate less. This type of data set is often called an *unevenly spaced time series*. However, for many types of environmental time series, observations are available at *equally spaced discrete time intervals* such as hourly, daily, weekly, monthly or yearly time separations. Average weekly precipitation records may be convenient for use in forecasting short-term weather trends while mean yearly records may be appropriate for studying longer-term climatic changes. In Parts II to IX of this book, as well as Chapter 22, the *time series models* considered are designed for use with discrete time series that are measured at equally spaced time intervals. Additionally, the variable being observed at discrete times is assumed to be measured as a continuous variable using the real number scale. Furthermore, the type of model to be employed is not only a function of the inherent properties of the phenomenon that is being modelled but is also dependent upon the time interval under consideration. For example, the *stationary nonseasonal models* of Chapter 3 are designed for fitting to average yearly riverflow series while the *seasonal models* of Chapters 13 and 14 can be used with average monthly riverflow time series. Finally, the *nonparametric trend tests* of Chapter 23, the *regression analysis models* of Chapter 24, and many of the *graphical methods* of Chapter 22 and elsewhere in the book, can be employed with both evenly and unevenly spaced measurements.

The assumption, that the entries in a time series under study are given at *discrete time intervals that are evenly spaced*, has many inherent advantages. Firstly, natural time series are often conveniently available in this type of format. Government agencies frequently list riverflows both as average weekly and monthly values. Other types of time series may only be given as one measurement during each time interval and, therefore, it is not possible to represent each entry in the time series as an average value. Secondly, the equispaced discrete time assumption greatly simplifies the mathematical theory underlying the various types of stochastic or time series models that can be designed for modelling environmental time series. In fact, little research has been successfully completed regarding comprehensive stochastic models that can allow for the time interval to vary between observations. Thirdly, if the data are not given in the form of an equally spaced discrete time series, the observations can often be conveniently converted to this format. Continuous time series can be easily transformed to discrete observations by lumping data together over a specified time interval. For instance, continuous temperature information may be listed as average hourly readings. Other types of data may be continuously accumulated over a period of time. For a chosen time interval, the amount accumulated over that period can form one value in the discrete time series. Rain gauges, for example, may be inspected weekly in order to record the amount of precipitation that has accumulated. In other situations, a discrete time series that is recorded using a specified time interval, may be changed to a data sequence that is based upon a larger time separation between observations. For instance, average daily riverflows can readily be converted to mean weekly, monthly or yearly records. In some situations, certain types of time series that do not possess equal time separations between observations may in fact be treated as if the time intervals were constant. For example, when the values in a time series represent the occurrence of some kind of event such as the successive yields from a batch chemical process, the amount of time that elapses between each happening may not be important. Consequently, the time series can be analyzed using the techniques that have been developed for equally spaced observations. Finally, as explained in Section 19.3 and elsewhere in the book, unevenly spaced series can often be converted to evenly spaced series by employing appropriate data filling procedures.

In most time series studies, the interval separating observations is time. However, it is possible to have other types of separations. The *interval may be spatial*. The depth of a lake at equally spaced intervals along its length may behave according to some probabilistic mechanism. The contours of a mountain range in a fixed direction could perhaps be treated as a time series. The values for the direction of flow of a meandering river measured at equispaced points along the course of the river, constitute a time series based upon spatial considerations (Speight, 1965; Ikeda and Parker, 1989). Nevertheless, in the vast majority of practical applications the spacing between observations in a series is due to time. Accordingly, even if the spacing between entries is a result of distance, the term "time" series is still usually employed.

If a polynomial can be fit to a known time series and future entries of the time series can be exactly determined, the time series is said to follow a *deterministic function*. When the future values of a time series cannot be calculated exactly and can be described solely in terms of a probability distribution, the time series is described by a *nondeterministic model* which is usually some kind of statistical or stochastic model. Chronological observations measured from a given phenomenon form a statistical time series. By knowing the historical values of the widths of the tree rings at a specified site, for example, the range of possible growths for the upcoming years can only be predicted using appropriate probabilistic statements. *This text is involved with modelling natural phenomena which evolve with time according to a probabilistic structure.*

2.3 STOCHASTIC PROCESS

For natural phenomena it is impossible to predict deterministically what will occur in the future. For instance, meteorologists never state that there will be exactly 3.00 mm of rain tomorrow. However, once an event, such as tomorrow's rainfall, has occurred, then that value of the precipitation time series is known exactly. Nevertheless, it will continue to rain in the future and the sequence of all the historical precipitation records is only one realization of what could have occurred and also of what could possibly happen. Precipitation is an example of a statistical phenomenon that evolves in time according to probabilistic laws. A mathematical expression which describes the probability structure of the time series that was observed due to the phenomenon, is referred to as a *stochastic process*. The sequence of historical observations is in fact a *sample realization* of the stochastic process that produced it.

In Table 1.4.1 within Section 1.4.3, stochastic models are classified according to the criteria of discrete and continuous time as well as discrete and continuous state space. As pointed out in Section 1.4.3, this book deals with time series models which constitute a special class of stochastic models for which the time is discrete and the possible values or state space of the variables being measured are continuous. Some well known books on stochastic processes include contributions by Cox and Miller (1965) referenced in Section 1.4.3, Parzen (1962), Ross (1983) and Papoulis (1984). Representative books on time series analysis are referred to in Section 1.6.3.

In a practical application, a time series model is fitted to a given series in order to calibrate the parameters of the model or stochastic process. The procedure of fitting a time series or stochastic model to the time series for use in applications is called *time series analysis*. One objective of time series analysis is to make inferences regarding the basic features of the stochastic process from the information contained in the historical time series. This can be accomplished by developing a mathematical model which possesses the same key statistical properties as the generating mechanism of the stochastic process, when the model is fit to the given time series. The fitted model can then be used for various applications such as forecasting and simulation. The families of stochastic models considered in this text constitute classes of processes that are amenable for modelling water resources and other natural time series.

In Part III, a linear nonseasonal model is designed for modelling the average annual flows of the St. Lawrence River at Ogdensburg, New York, U.S.A., from 1860 to 1957. The average flows are calculated in m^3/s for the water year from October 1 of one year to September 30 of the following year and were obtained from a paper by Yevjevich (1963). Figure 2.3.1 shows a plot of the 97 observations. As explained in Chapter 9, the model which is fitted to the flows can be used to generate or *simulate* other possible sequences of the flows. For instance, Figures 2.3.2 and 2.3.3 display two generated sequences from the fitted model. Notice that the synthetic time series shown in these two figures differ from each other and are also not the same as the historical series in Figure 2.3.1. However, within the confines of the fitted model the generated series do possess the same overall statistical characteristics of the historical data. In general, an *ensemble* of data sequences could be generated to portray a set of possible realizations from the population of time series that are defined by the generating stochastic process.

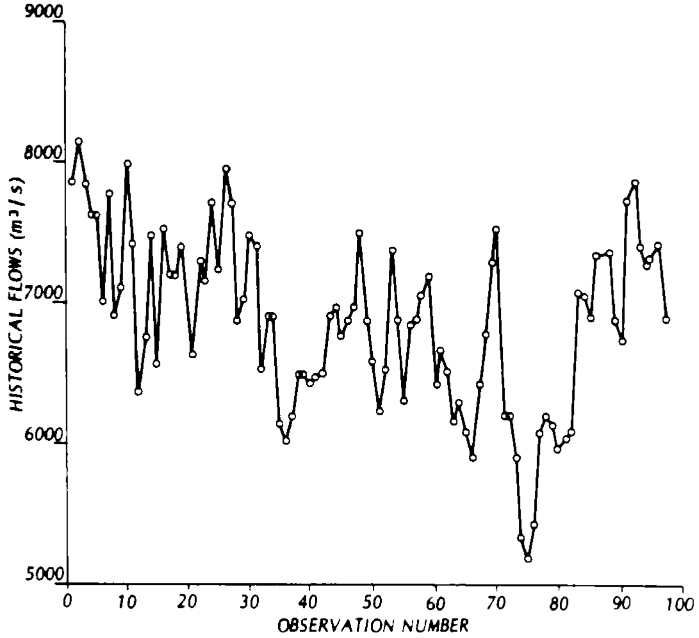


Figure 2.3.1. Annual flows of the St. Lawrence River at Ogdensburg, New York.

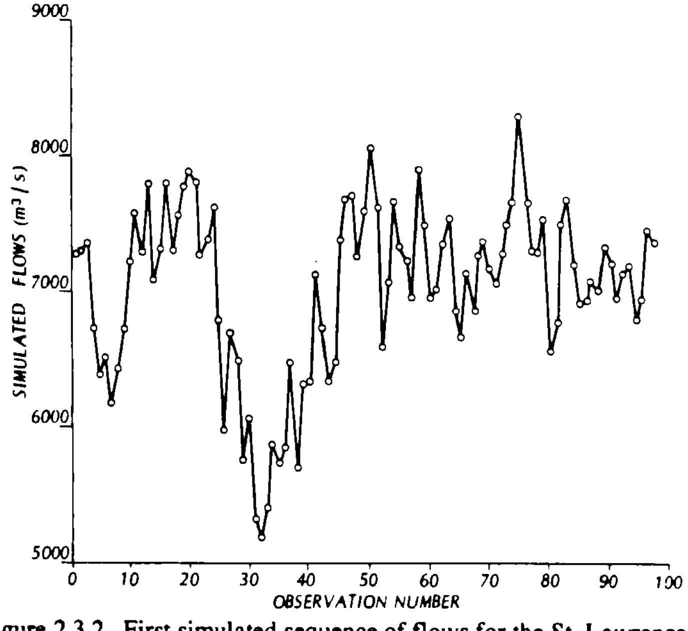


Figure 2.3.2. First simulated sequence of flows for the St. Lawrence River at Ogdensburg, New York.

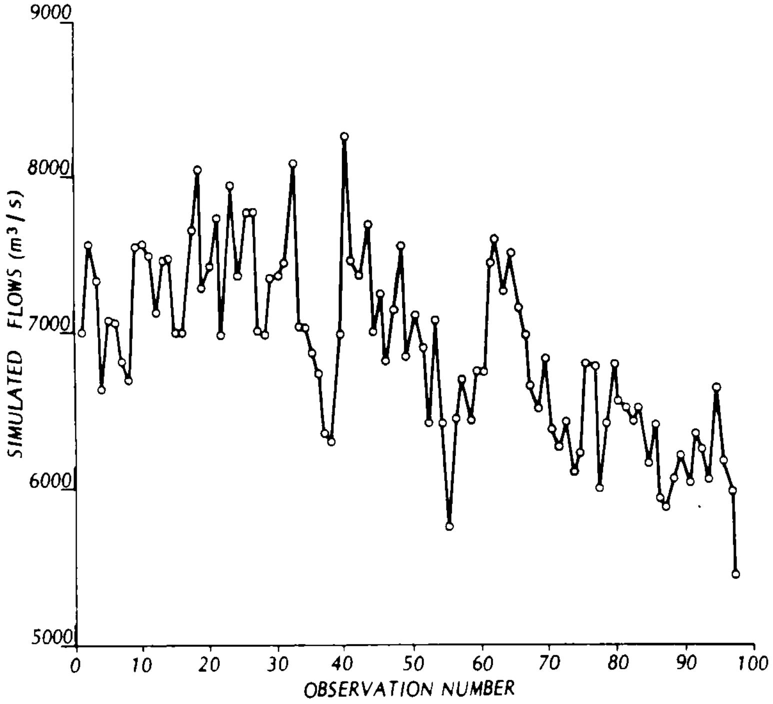


Figure 2.3.3. Second simulated sequence of flows for the St. Lawrence River at Ogdensburg, New York.

Because it is conceptually possible for more than one sequence of values to occur over a specified time span, a stochastic process can theoretically be represented by a *random variable* at each point in time. Each random variable possesses its own marginal probability distribution while joint probability distributions describe the probability characteristics of more than one random variable. In order to simplify the mathematical theory underlying a stochastic process, it is often assumed that the stochastic process is stationary.

2.4 STATIONARITY

2.4.1 General Discussion

Stationarity of a stochastic process can be qualitatively interpreted as a form of statistical equilibrium. Therefore, the statistical properties of the process are not a function of time. For example, except for inherent stochastic fluctuations, stationary stochastic models are usually designed such that the mean level and variance are independent of time. Besides reducing the mathematical complexity of a stochastic model, the stationarity assumption may reflect reality. For instance, if a natural river basin has not been subjected to any major land use changes such as urbanization and cultivation, it may be reasonable to assume that a stationary stochastic model can be fitted to the time series of historical average annual riverflows. Consequently, this infers that the stochastic properties of the complex physical mechanism that produces the observed riverflows, can be represented mathematically by a stationary stochastic process.

Stationarity is analogous to the concept of *isotropy* within the field of physics. In order to be able to derive physical laws that are deterministic, it is often assumed that the physical properties of a substance such as conductivity and elasticity, are the same regardless of the direction or location of measurement. For example, when studying the conductive properties of an electrical transmission line, it is reasonable to consider the wire to have uniform cross-sectional area and constant density of copper along its length. Likewise, in stochastic modelling, the statistical properties of a process are invariant with time if the process is stationary.

In certain situations, the statistical characteristics of a process are a function of time. Water demand tends to increase over the years as metropolitan areas grow in size and the affluence of the individual citizen expands. The average carbon dioxide content of the atmosphere may increase with time due to complex natural processes and industrial activities. To model an observed time series that possesses *nonstationarity*, a common procedure is to first remove the nonstationarity by invoking a suitable transformation and then to fit a stationary stochastic model to the transformed sequence. For instance, as explained in Section 4.3.1, one method to remove nonstationarity is to difference the given data before determining an appropriate stationary model. Therefore, even when modelling nonstationary data, the mathematical results that are available for describing stationary processes, are often required.

The idea of stationarity is a mathematical construct that was created to simplify the theoretical and practical development of stochastic models. Even the concept of a stochastic process was adopted for mathematical convenience. For a particular geophysical or other type of natural phenomenon, the only thing that is actually known is one unique historical series. An ensemble of possible time series does not exist because the clocks of nature cannot be turned back in order to produce more possible time series. Consequently, Klemes (1974, p. 676) maintains that it is an exercise in futility to argue on mathematical grounds about the stationarity or nonstationarity of a specific geophysical series. Rather, the question of whether or not a process is stationary is probably a philosophical one and is based upon an understanding of the system being studied.

Some researchers believe that natural processes are inherently nonstationary and therefore the greater the time span of the historical series, the greater is the probability that the series will exhibit statistical characteristics which change with time. However, for relatively short time spans it may be feasible to approximately model the given data sequence using a stationary stochastic model. Nevertheless, the reverse position may seem just as plausible to other scientists. Apparent nonstationarity in a given time series may constitute only a local fluctuation of a process that is in fact stationary on a longer time scale.

Within this textbook, *the question of stationary or its antipode, is based upon practical considerations*. When dealing with yearly hydrological and other kinds of natural time series of moderate time spans, it is often reasonable to assume that the process is approximately stationary (Yevjevich, 1972a,b). For example, even though the climate may change slowly over thousands of years, within the time span of a few hundred years the changes in hydrologic time series may be relatively small and therefore these series can be considered to be more or less stationary. If the underlying modelling assumptions are satisfied when a stationary stochastic model is fitted to a nonseasonal series, then these facts validate the assumption of stationarity. When considering average monthly riverflows, the individual monthly averages may have constant mean values but the means may vary from month to month. Therefore, as explained in Chapters 13 and 14, time series models are employed that reflect the stationarity properties within a given month but recognize the nonstationarity characteristics across all of the months. In other situations, there may be a physical reason for a process to undergo a change in mean level. For example, in 1961 a forest fire in Newfoundland, Canada, devastated the Pipers Hole River basin. In Section 19.5.4, an intervention model is used to model the monthly flows of the Pipers Hole River before and after the fire. The intervention model describes the manner in which the riverflows return to their former patterns as the natural vegetation slowly reverts, over the years, to its condition prior to the fire.

2.4.2 Types of Stationarity

As mentioned previously, the historical time series can be thought of as one realization of the underlying stochastic process that generated it. Consequently, a stochastic process can be represented by a random variable at each point in time. When the joint distribution of any possible set of random variables from the process is unaffected by shifting the set backwards or forwards in time (i.e., the joint distribution is time independent), then the stochastic process is said to possess *strong (or strict) stationarity.*

In practice, the assumption of strong stationarity is not always necessary and a weaker form of stationarity can be assumed. When the statistical moments of the given time series up to order k depend only on time differences and not upon the time of occurrence of the data being used to estimate the moments, the process has *weak stationarity* of order k. For example, if the stochastic process can be described by its mean, variance and autocorrelation function (ACF) (see Section 2.5.2 for the definition of the ACF), then it has *second-order stationarity.* This second-order stationarity may also be referred to as *covariance stationarity* and all of the stationary processes discussed in this text are covariance stationary. Some important statistics which are used in conjunction with covariance stationary processes, are now defined.

2.5 STATISTICAL DEFINITIONS

In this section, some basic definitions are presented that are especially useful in the time series analysis. Readers who have forgotten some of the basic ideas in probability and statistics are encouraged to refresh their memories by referring to some introductory books such as the ones by Ross (1987), Kalbfleisch (1985), Snedecor and Cochran (1980), Kempthorne and Folks (1971) and Guttman et al. (1971) as well as statistical hydrology books by writers including McCuen and Snyder (1986), Haan (1977) and Yevjevich (1972a). Moreover, a handbook on statistics is provided by Sachs (1984) while Kotz and Johnson (1988) are editors of a comprehensive encyclopedia on statistics.

2.5.1 Mean and Variance

Let $z_1,z_2, \cdots ,z_N$, be a time series of N values that are observed at equispaced time intervals. The theoretical *mean* $\mu = E[z_t]$ of the process can be estimated from the sample realization by using the equation

$$\bar{z} = \frac{1}{N}\sum_{t=1}^{N} z_t \qquad [2.5.1]$$

The amount of spread of a process about its mean μ is related to its theoretical *variance* $\sigma_z^2 = E[(z_t - \mu)^2]$. This variance can be estimated from the given time series by employing the equation

$$\hat{\sigma}_z^2 = \frac{1}{N}\sum_{t=1}^{N}(z_t - \bar{z})^2 \qquad [2.5.2]$$

2.5.2 Autocovariance and Autocorrelation

The *covariance* between z_t and a value z_{t+k} which is k time lags removed from z_t, is theoretically defined in terms of the *autocovariance* at lag k given by

$$\gamma_k = cov[z_t, z_{t+k}] = E[(z_t - \mu)(z_{t+k} - \mu)] \qquad [2.5.3]$$

When k=0, the autocovariance is the variance and consequently $\gamma_o = \sigma_z^2$.

A normalized quantity that is more convenient to deal with than γ_k, is the theoretical *autocorrelation coefficient* which is defined at lag k as

$$\rho_k = \frac{\gamma_k}{\gamma_o} \qquad [2.5.4]$$

Because of the form of [2.5.4], the autocorrelation coefficient is dimensionless and, therefore, independent of the scale of measurement. Furthermore, the possible values of ρ_k range from -1 to 1, where ρ_k has a magnitude of unity at lag zero.

Jenkins and Watts (1968, p. 146) refer to the autocovariance, γ_k, as the theoretical *autocovariance function* while the autocorrelation coefficient, ρ_k, is called the theoretical *autocorrelation function (ACF)*. Although the ACF is also commonly referred to as the autocorrelation coefficient or *serial correlation coefficient*, in this book the terminology ACF is employed. For interpretation purposes, it is often useful to plot the ACF against lag k. Because the ACF is symmetric about lag zero, it is only necessary to plot ρ_k for positive lags from lag one onwards.

Autocovariance and Autocorrelation Matrices

Let the N historical observations be contained in the vector

$$\mathbf{z}^T = (z_1, z_2, \cdots, z_N)$$

The *autocovariance matrix* for a stationary process of N successive observations is defined by

$$\Gamma_N = E[(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^T]$$

where $\boldsymbol{\mu}$ is a vector of dimension N×1 which contains N identical entries for the theoretical mean level μ. In expanded form, the autocovariance matrix is a doubly symmetric matrix and is written as

$$\Gamma_N = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{N-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{N-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{N-3} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \gamma_{N-1} & \gamma_{N-2} & \gamma_{N-3} & \cdots & \gamma_0 \end{bmatrix} \qquad [2.5.5]$$

The *autocorrelation matrix* is defined by

$$\mathbf{P}_N = \frac{\Gamma_N}{\gamma_0} \qquad [2.5.6]$$

For the random variables, $z_t, z_{t-1}, \cdots, z_{t-N+1}$, consider any linear function given by

$$L_t = l_1(z_t - \mu) + l_2(z_{t-1} - \mu) + \cdots + l_N(z_{t-N+1} - \mu)$$

By letting $\mathbf{l}$ be the vector $\mathbf{l}^T = (l_1, l_2, \cdots, l_N)$, the linear function can be economically written as $L_t = \mathbf{l}^T(\mathbf{z} - \mu)$. For a stationary process, the covariance function is symmetric about lag zero and hence $cov[z_i z_j] = \gamma_{|j-i|}$. Consequently, the variance of L_t is

$$var[L_t] = cov[L_t L_t] = E[L_t L_t^T] = E[\mathbf{l}^T(\mathbf{z} - \mu)(\mathbf{l}^T(\mathbf{z} - \mu))^T] = E[\mathbf{l}^T(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T \mathbf{l}]$$
$$= \mathbf{l}^T E[(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T]\mathbf{l} = \mathbf{l}^T \Gamma_N \mathbf{l} = \sum_{i=1}^{N}\sum_{j=1}^{N} l_i l_j \gamma_{|j-i|}$$

If the l's are not all zero and the series is nondeterministic, then $var[L_t]$ is strictly greater than zero and hence the quadratic form in the above equation is positive definite. Therefore, it follows that the autocovariance and autocorrelation matrices are positive definite for any stationary process (Box and Jenkins, 1976, p. 29). Consequently, the determinant and all the principal minors of these matrices must be greater than zero.

When the probability distribution associated with a stochastic process is a multivariate normal distribution, then the process is said to be a normal or a *Gaussian process*. Because the multivariate normal distribution is completely characterized in terms of the moments of first and second order, the presence of a mean and autocovariance matrix Γ_N for all N implies that the process possesses strict stationarity. In addition, when the process is Gaussian, the ACF completely characterizes all of the dependence in the series.

2.5.3 Short and Long Memory Processes

For a known stochastic process, it is possible to determine the theoretical autocovariance, γ_k, or equivalently the theoretical ACF, ρ_k. In Chapter 3, for example, theoretical ACF's are derived for different kinds of stationary autoregressive-moving average (ARMA) processes, while in Section 10.4 the theoretical ACF is presented for a fractional Gaussian noise (FGN) process. When the theoretical ACF is *summable* it must satisfy (Brillinger, 1975)

$$M = \sum_{k=-\infty}^{\infty} |\rho_k| < \infty \qquad [2.5.7]$$

where M stands for memory. Essentially, a covariance stationary process is said to possess a *short memory* or *long memory* according to whether or not the theoretical ACF is summable. For more precise definitions of short and long memory, the reader can refer to Cox (1991). Examples of short memory processes are the stationary ARMA processes in Chapter 3 whereas the FGN and fractional ARMA (FARMA) processes of Chapters 10 and 11, respectively, possess long memory for specified ranges of certain model parameters. The importance of both long and short memory processes for modelling annual hydrological time series is exemplified by the study of the "Hurst phenomenon" in Chapter 10.

2.5.4 The Sample Autocovariance and Autocorrelation Functions

In practical applications, the autocovariance function and the ACF are estimated from the known time series. Jenkins and Watts (1968) have studied various procedures for estimating the autocovariance function from the given sample of data. It is concluded that the most appropriate sample estimate of γ_k, the autocovariance at lag k, is

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z}) \qquad [2.5.8]$$

The estimated or *sample ACF* for kth lag autocorrelation ρ_k is

$$r_k = \frac{c_k}{c_o}. \qquad [2.5.9]$$

To obtain the *sample autocovariance matrix*, one substitutes c_k from [2.5.8] for γ_k, $k = 0,1,\ldots,N-1$, into [2.5.5]. Using the divisor N in [2.5.8] instead of $N-k$ insures that the sample autocovariance matrix is positive definite (McLeod and Jimenez, 1984). Because the sample autocovariance matrix is positive definite for a stationary process, this property also holds for the sample autocovariance matrix as well as the sample ACF matrix.

As explained for the case of ARMA models in Chapters 3 and 5, the sample ACF is useful for identifying what type of time series model to fit to a given time series of length N. Because the ACF is symmetric about lag zero, it is only required to plot the sample ACF for positive lags except for lag zero, to a maximum lag of about $N/4$. To determine which values of the estimated ACF are significantly different from zero, confidence limits should also be included on the graph. This requires a knowledge of the variance of the sample ACF, r_k.

For short-memory processes, the *approximate variance* for r_k is given by Bartlett (1946) as

$$var[r_k] \approx \frac{1}{N}\sum_{j=-\infty}^{+\infty}(\rho_j^2 + \rho_{j+k}\rho_{j-k} - 4\rho_k\rho_j\rho_{j-k} + 2\rho_j^2\rho_k^2) \qquad [2.5.10]$$

The above equation can be greatly simplified if it is known that ρ_j is zero beyond lag q. In particular, the variance of r_k after lag q is derived from [2.5.10] as

$$var[r_k] \approx \frac{1}{N}(1 + 2\sum_{j=1}^{q}\rho_j^2) \qquad \text{for } k>q \qquad [2.5.11]$$

When a normal process is uncorrelated and $\rho_k = 0$ for $k > 0$, the variance of r_k for $k > 0$ is approximately $\frac{1}{N}$ from [2.5.11]. Using simulation experiments, Cox (1966) demonstrated that when r_1 is calculated for a sequence of uncorrelated samples, the sampling distribution of r_1 is very stable under changes of distribution and the asymptotic normal form of the sampling distribution is a reasonable approximation even in samples as small as ten. However, for correlated data larger samples are required in order for [2.5.11] to be valid.

When using [2.5.11] in practice, the first step is to substitute r_k for $\rho_k(k = 1,2, \cdots ,q)$ into the equation if ρ_k is assumed to be zero after lag q. Then, the square root of the estimated variance for r_k can be calculated to determine the large-lag estimated standard deviation. An estimated standard deviation, such as the one just described, is commonly referred to as a *standard error (SE)*. Moreover, because the distribution of r_k is approximately normal, appropriate confidence limits can be established. For instance, to obtain the 95% confidence interval (or equivalently the 5% significance interval) at a given lag, plot 1.96 times the large-lag SE above and below the axis. When determining the sample ACF, one has the option of either estimating the mean of the input series when employing [2.5.9] to calculate the sample ACF or else assuming the mean to be zero. If one is examining the sample ACF of the given series, the mean should be estimated for use in [2.5.9]. If it is found that the data are not stationary, the nonstationary can sometimes be removed by an operation called differencing (see Section 4.3.1). The mean of series that remains after differencing is usually zero (refer to [4.3.2]) and, consequently, when estimating the ACF for such a series the mean can be set equal to zero. If it is suspected that there is a deterministic trend component contained in the data, the mean of the differenced series should be removed when estimating the ACF for the differenced series (see Section 4.6). Finally, the mean is assumed to be zero for the sequence of residuals that can be estimated when a linear time series model is fitted to a specified data set. Therefore, when calculating the residual ACF, a mean of zero is employed (see Section 7.3).

Average annual temperature data are available in degrees Celsius for the English Midlands from 1813-1912 (Manley, 1953, pp. 225-260). Equations [2.5.8] and [2.5.9] are employed to calculate r_k while the 95% confidence limits are obtained using [2.5.11] if it is assumed that ρ_k is zero after lag q. Figure 2.5.1 is a plot of the estimated ACF for the temperature data. Notice that there are rather large values of the ACF at lags 1, 2 and 15. Because the data are nonseasonal, the magnitude of the sample ACF at lag 15 could be due to chance. When ρ_k is assumed to be zero after lag 2, the 95% confidence limits of the sample ACF for the temperature data are as shown in Figure 2.5.2.

The theoretical ACF can also be plotted for the temperature data. After fitting a proper stationary ARMA model to these data (see Section 3.3.2 and Part III), the known parameter estimates can be utilized to calculate the theoretical ACF (see Sections 3.3.2 and 3.4.2, and Appendix A3.2 for theoretical descriptions). The theoretical ACF for the temperature data is displayed in Figure 2.5.3. Notice that the plots given in Figures 2.5.2 and 2.5.3 are very similar. As is explained in Chapter 10, when an appropriate time series model is properly fitted to a given data set, the fitted model will preserve the important historical statistics such as the sample ACF at

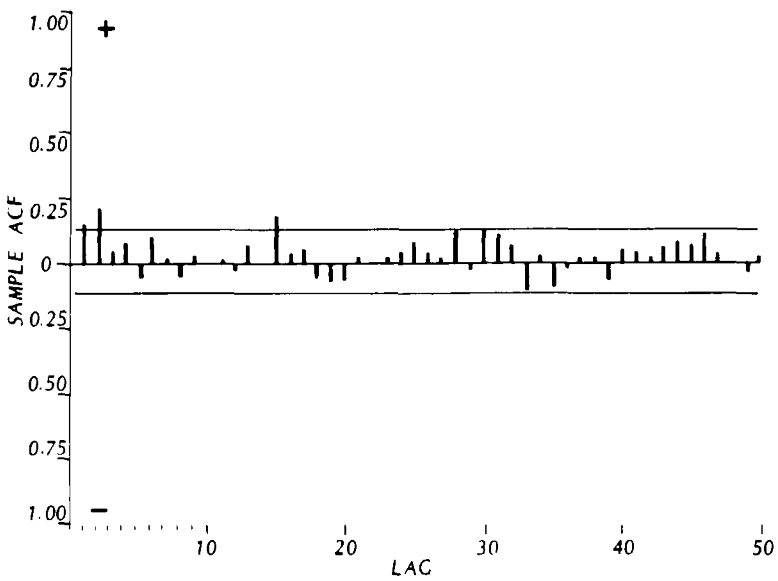


Figure 2.5.1. Sample ACF and 95% confidence limits for the annual temperature data from the English Midlands.

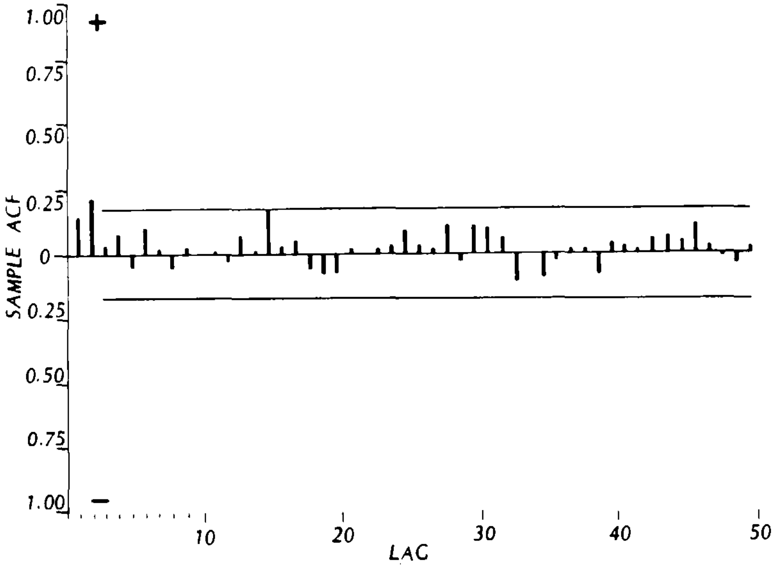


Figure 2.5.2. Sample ACF and 95% confidence limits for the annual temperature data from the English Midlands when ρ_k is zero after lag 2.

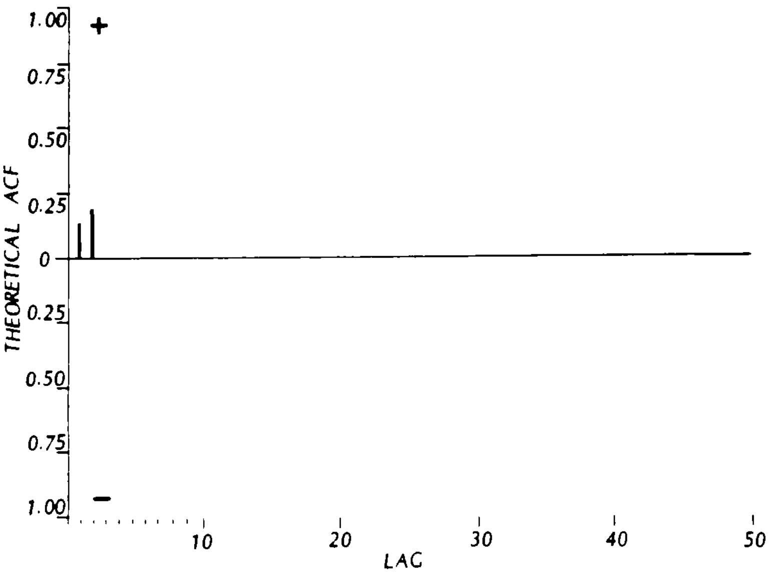


Figure 2.5.3. Theoretical ACF for the model fitted to the temperature data from the English Midlands.

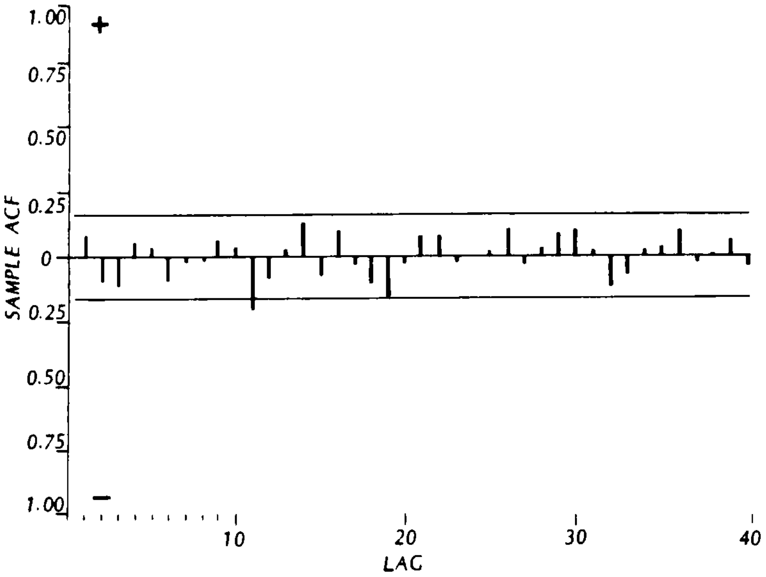


Figure 2.5.4. Sample ACF and 95% confidence limits for the average annual flows of the Rhine River at Basle, Switzerland.

different lags. It is crucial that stochastic models that are used in practice possess a theoretical ACF that is close to the sample ACF, especially at lower lags.

The observations in many yearly hydrological data are often uncorrelated. Consider the average annual flows of the Rhine River in m^3/s at Basle, Switzerland. These flows are given from 1837 to 1957 in a paper by Yevjevich (1963). As shown by the sample ACF in Figure 2.5.4, the Rhine flows appear to be uncorrelated except for a value of lag 11 which could be due to chance alone. The 95% confidences limits are calculated using [2.5.11], under the assumption that ρ_k is zero for all nonzero lags.

The plot of the theoretical ACF of the Rhine flows would be exactly zero at all nonzero lags. The observations are, therefore, uncorrelated and are called white noise (see discussion on spectral analysis in Section 2.6 for a definition of white noise). If the time series values are uncorrelated and follow a multivariate normal distribution, the white noise property implies independence. When the observations are not normal, then lack of correlation does not necessarily infer independence. However, independence always means that the observations are uncorrelated.

Some care must be taken when interpreting a graph of the sample ACF. Bartlett (1946) has derived formulae for approximately calculating the covariances between two estimates of ρ_k at different lags. For example, the *large lag approximation for the covariance between r_k and r_{k+i}* assuming $\rho_j = 0$ for $j \geq k$ is

$$cov[r_k, r_{k+i}] \simeq \frac{1}{N} \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+i} \qquad [2.5.12]$$

An examination of [2.5.12] reveals that large correlations can exist between neighbouring values of r_k and can cause spurious patterns to appear in the plots of the sample ACF.

2.5.5 Ergodicity Conditions

A desirable property of an estimator is that as the sample size increases the estimator converges with probability one to the population parameter being estimated. An estimator possessing this property is called a *consistent estimator*. To estimate the mean, variance and ACF for a single time series, formulae are presented in [2.5.1], [2.5.2] and [2.5.9], respectively. In order for these estimators to be *consistent*, the stochastic process must possess what is called *ergodicity*. Another way to state this is that an ensemble statistic, such as the mean, across all possible realizations of the process at each point in time, is the same as the sample statistic for the single time series of observations. For a detailed mathematical description of ergodicity, the reader may wish to refer to advanced books in stochastic processes [see for example Hannan (1970, p. 201), Parzen (1962, pp. 72-76), and Papoulis (1984, pp. 245-254)].

For a process, z_t, to be mean-ergodic and the sample mean $\bar{z}$ in [2.5.1] constitute a consistent estimator for the theoretical mean μ, a necessary and sufficient condition is

$$\lim_{N \to \infty} var(\bar{z}_N) = 0 \qquad [2.5.13]$$

where $\bar{z}_N$ is the sample mean of a series having N observations. Sufficient conditions for mean-ergodocity are:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N}\gamma_k=0 \qquad [2.5.14]$$

or $cov(z_t,\bar{z}_N)\to 0$ as $N\to\infty$ or $\gamma_k\to 0$ as $k\to\infty$

A process represented by z_t is said to be *Gaussian* if any linear combination of the process is normally distributed. When the process is Gaussian, a sufficient condition for ergodicity of the autocovariance function is the theoretical autocovariances in [2.5.8] satisfy

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N}\gamma_k^2=0 \qquad [2.5.15]$$

From the above formulae, it can be seen that ergodicity implies that the autocovariance or autocorrelation structure of the time series must be such that the present does not depend "too strongly" on the past. All stationary time series models that are used in practice have ergodic properties.

2.6 SPECTRAL ANALYSIS

The *spectrum* is the Fourier transform of the autocovariance function (Jenkins and Watts, 1968) and, therefore, provides no new information about the data that is not already contained in the autocovariance function or equivalently the ACF. However, the spectrum does provide a different interpretation of the statistical properties of the time series since it gives the distribution of the variance of the series with frequency. As shown by Jenkins and Watts (1968), the spectrum can be plotted against frequency in the range from 0 to 1/2. Therefore, when studying the spectrum one is said to be working in the *frequency domain* while investigating the autocovariance function or ACF is referred to as studying in the *time domain*. For the topics covered in this text, it is usually most convenient to carry out time series studies in the time domain. Nevertheless, occasionally a spectral analysis can furnish valuable insight in certain situations. In Section 3.5, the theoretical spectra of ARMA processes are presented. The cumulative periodogram, which is closely related to the cumulative spectrum, can be utilized at the identification and diagnostic check stages of model development (see Part III). Due to its usefulness in forthcoming topics within the book, the cumulative periodogram is now described.

Given a stationary time series $z_1,z_2,\cdots,z_N$, the periodogram function, $I(f_j)$, is

$$I(f_j)=\frac{2}{N}\left|\sum_{t=1}^{N}z_t\exp(-2\pi i f_j t)\right|$$
$$=\frac{2}{N}\left[\left(\sum_{t=1}^{N}z_t\cos 2\pi f_j t\right)^2+\left(\sum_{t=1}^{N}z_t\sin 2\pi f_j t\right)^2\right]^{\frac{1}{2}} \qquad [2.6.1]$$

where $f_j=\frac{j}{N}$ is the jth frequency $j=1,2,\ldots,N'$, $N'=[N/2]$ (take integer portion of $N/2$), $|\cdot|$ denotes the magnitude and $i=\sqrt{-1}$. In essence, $I(f_j)$ measures the strength of the relationship between the data sequence z_t and a sinusoid with frequency f_j where $0<f_j\leq 0.5$.

The normalized *cumulative periodogram* is defined by

$$C(f_k) = \frac{\sum_{j=1}^{k} I(f_j)}{N\hat{\sigma}_z^2} \qquad [2.6.2]$$

where $\hat{\sigma}_z^2$ is the estimated variance defined in [2.5.2]. The normalized cumulative periodogram is henceforth simply referred to as the cumulative periodogram.

When estimating the cumulative periodogram, sine and cosine terms are required in the summation components in [2.6.1]. To economize on computer usage, the sum-of-angles method can be used to recursively calculate the sine and cosine terms by employing (Robinson, 1967, p. 64; Otnes and Enochson, 1972, p. 139)

$$\cos 2\pi f_j(t+1) = a\cos 2\pi f_j t - b\sin 2\pi f_j t \qquad [2.6.3]$$

$$\sin 2\pi f_j(t+1) = b\cos 2\pi f_j t + a\sin 2\pi f_j t \qquad [2.6.4]$$

where

$$a = \cos 2\pi f_j \text{ and } b = \sin 2\pi f_j$$

Utilization of the above relationships does not require any additional computer storage and is much faster than using a standard library function to evaluate repeatedly the sine and cosine functions.

When $C(f_k)$ is plotted against f_k, the ordinate $C(f_k)$ ranges from 0 to 1 while the abscissa f_k goes from 0 to 0.5. Note that

$$\sum_{j=1}^{N'} I(f_j) = N\hat{\sigma}_z^2$$

Therefore, if the series under consideration were uncorrelated or *white noise*, then a plot of the cumulative periodogram would consist of a straight line joining (0,0) and (0.5,1). The term white noise is employed for an uncorrelated series, since the spectrum of such a series would be evenly distributed over frequency. This is analogous to white light which consists of electromagnetic contributions from all of the visible light frequencies.

In order to use the cumulative periodogram to test for white noise, confidence limits for white noise must be drawn on the cumulative periodogram plot parallel to the line from (0,0) to (0.5,1). For an uncorrelated series, these limits would be crossed a proportion ε of the time. The limits are drawn at vertical distances $\pm\frac{K_\varepsilon}{\sqrt{\left[\frac{N-1}{2}\right]}}$ above and below the theoretical white noise line, where $\left[\frac{N-1}{2}\right]$ means to take only the integer portion of the number inside the brackets. Some approximate values for K_ε are listed in Table 2.6.1.

Table 2.6.1. Parameters for calculating confidence limits for the cumulative periodogram.

ε	0.01	0.05	0.10	0.25
K_ε	1.63	1.36	1.22	1.02

Unlike spectral estimation, the cumulative periodogram white noise test is useful even when only a small sample (at least 50) is used. The cumulative periodogram for the average annual flows of the Rhine River at Basle, Switzerland from 1937-1957 is given in Figure 2.6.1. As shown in this figure, the values for cumulative periodogram for the Rhine flows do not deviate significantly from the white noise line and fail to cross the 95% confidence limits. However, as illustrated by the cumulative periodogram in Figure 2.6.2, the average annual temperature data for the English Midlands from 1813-1912 are not white noise since the cumulative periodogram goes outside of the 95% confidence limits.

Besides being employed to test for whiteness of a given time series or perhaps the residuals of a model fitted to a data set, the cumulate periodogram has other uses. It may be used to detect hidden periodicities in a data sequence or to confirm the presence of suspected periodicities. For instance, annual sunspot numbers are available from 1700 to 1960 (Waldmeier, 1961) and the cumulative periodogram for the series is shown in Figure 2.6.3. Granger (1957) found that the periodicity of sunspot data follows a uniform distribution with a mean of about 11 years. This fact is confirmed by the dramatic jump in the cumulative periodogram where it cuts through the 95% confidence limits at a frequency of about $\frac{1}{11} = 0.09$.

Monthly riverflow data follow a seasonal cycle due to the yearly rotation of the earth about the sun. Average monthly riverflow data are available in m^3/s for the Saugeen River at Walkerton, Ontario, Canada, from January 1915 until December 1976 (Environment Canada, 1977) Besides the presence of a sinusoidal or cyclic pattern in a plot of the series against time, the behaviour of the cumulative periodogram can also be examined to detect seasonality. Notice in Figure 2.6.4 for the cumulative periodogram of the Saugeen River flows, that the cumulative periodogram cuts the 95% confidence limits at a frequency of 1/12 and spikes occur at other frequencies which are integer multiples of 1/12. Thus, seasonality is easily detected by the cumulative periodogram. In other instances, the cumulative periodogram may reveal that seasonality is still present in the residuals of a seasonal model that is fit to the data. This could mean that more seasonal parameters should be incorporated into the model to cause the residuals to be white noise (see Part VI).

2.7 LINEAR STOCHASTIC MODELS

This text is concerned mainly with linear stochastic models for fitting to stationary time series (see, for example, Chapter 3). When dealing with nonstationary data, stationary linear stochastic models can also be employed. By utilizing a suitable transformation, nonstationarity (such as trends, seasonality and variances changes over time) is first removed and then a linear stochastic model is fitted to the resulting stationary time series (see, for instance, Section 4.3). The usefulness and importance of linear stochastic models for modelling stationary time series is emphasized by the *Wold decomposition theorem.*

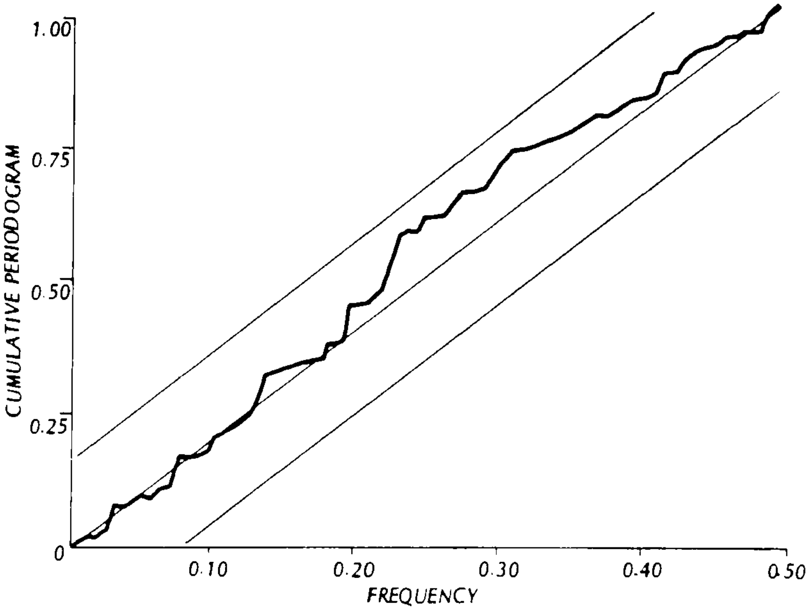


Figure 2.6.1. Cumulative periodogram and 95% confidence limits for the annual Rhine River flows at Basle, Switzerland.

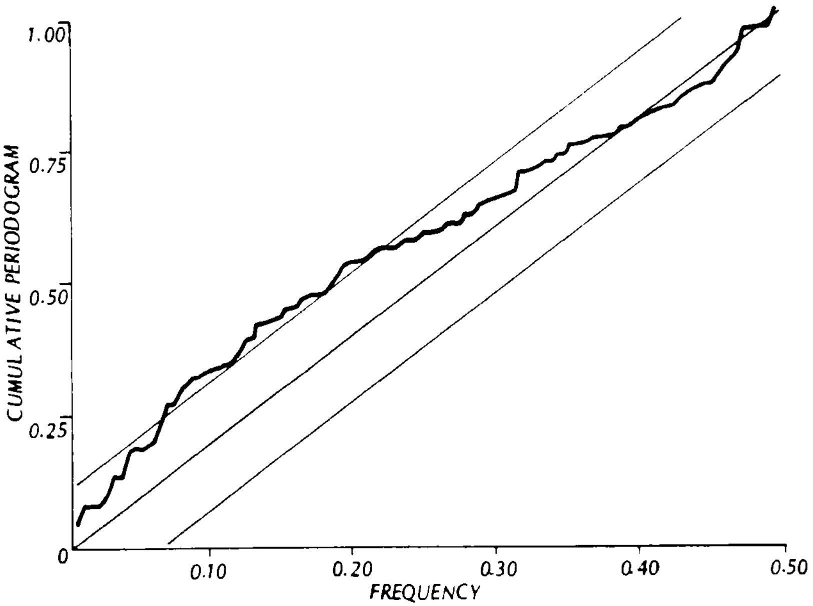


Figure 2.6.2. Cumulative periodogram and 95% confidence limits for the annual temperature data from the English Midlands.

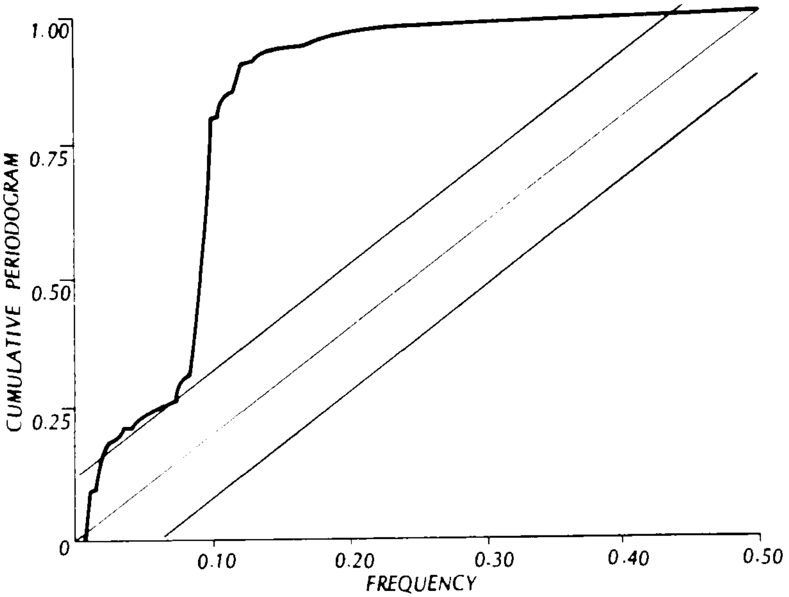


Figure 2.6.3. Cumulative periodogram and 95% confidence limits for the annual sunspot numbers from 1700 to 1960.

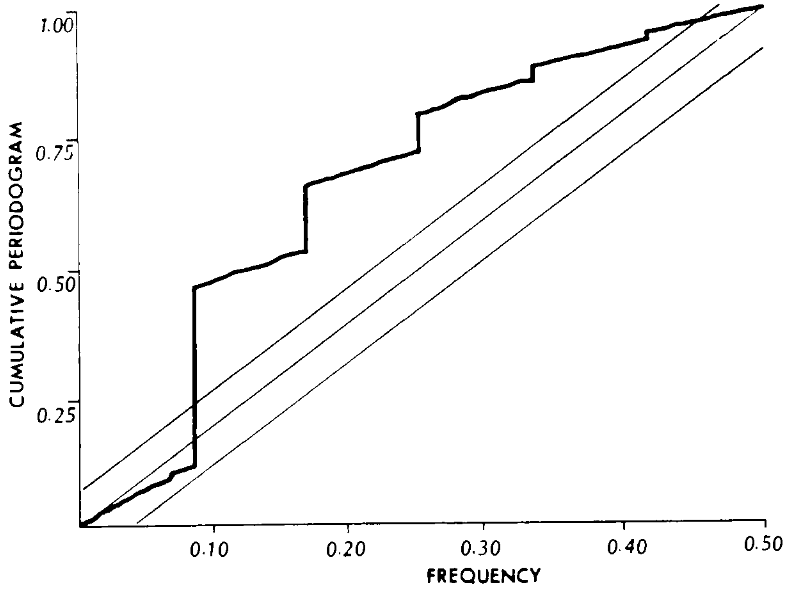


Figure 2.6.4. Cumulative periodogram and 95% confidence limits for the average monthly flows of the Saugeen River.

Wold (1954) proved that any stationary process, z_t, can be represented as the sum of a deterministic component and an uncorrelated purely nondeterministic component. The process for z_t at time t can be written as

$$z_t = \mu_t + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \qquad [2.7.1]$$

where t is discrete time that occurs at equispaced time intervals, μ_t is the deterministic component, a_t is white noise (also called disturbance, random shock or innovation) at time t, and ψ_i is the ith moving average parameter for which $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ for stationarity. The white noise, a_t, has the properties

$$E(a_t) = 0$$

$$var(a_t) = \sigma_a^2$$

and

$$cov(a_t, a_s) = 0, \quad t \neq s$$

The deterministic component, μ_t, can be a function of time or may be a constant such as the mean level μ of a process. The terms other than μ_t on the right hand side of [2.7.1] form what is called an infinite moving average (MA) process (see Section 3.4.3).

When a time series represented by z_t is Gaussian, the a_t's in [2.7.1] are independent and normally distributed with a mean of zero and a variance of σ_a^2. Consequently, the Wold decomposition theorem justifies the use of linear stochastic models for fitting to Gaussian stationary time series. In Part III, it is shown that many types of annual geophysical time series appear to be approximately Gaussian and stationary, and hence can be readily modelled using linear stochastic models. Furthermore, when the data are not normally distributed and perhaps also nonlinear, a Box-Cox transformation (Box and Cox, 1964) can be invoked to cause the transformed data to be approximately Gaussian and linear. Following this, a linear stochastic model can be fitted to the transformed series (see Section 3.4.5).

As is discussed in Section 3.4, the ARMA family of linear time series models constitutes a parsimonious representation of the infinite MA component that is given in [2.7.1]. The infinite number of MA parameters can be economically represented by a finite number (usually not more than four) of model parameters. Thus, the ARMA family of linear stochastic models are of utmost importance in time series modelling.

There is a close analogy between the Wold decomposition theorem and an important property from multiple linear regression. In the linear regression of the dependent variable y on the m independent variables $x_1, x_2, \ldots, x_m$, the error is uncorrelated with $x_1, x_2, \ldots, x_m$. For a stationary time series regression of z_t on its infinite past $z_{t-1}, z_{t-2}, \ldots,$ the a_t errors are uncorrelated with $z_{t-1}, z_{t-2}, \ldots$. Additionally, the a_t's are white noise.

As pointed out by Yule (1927), the a_t disturbances are fundamentally different from the superimposed type of error in other types of statistical models. This is because the a_t sequence in [2.7.1] affects not only the current observation, z_t, but the future values, $z_{t+1}, z_{t+2}, \cdots$, as well.

Consequently, the system is driven by the a_t innovations.

2.8 CONCLUSIONS

A covariance stationary time series can often be usefully described by its mean, variance and sample ACF or, equivalently, by its mean, variance and spectrum. For the types of applications considered in this book, it is usually most convenient to work in the time domain rather than the frequency domain. However, the cumulative periodogram is one of the concepts from spectral analysis that is used in some applications presented in the book.

Historically, the mean, variance and ACF have formed the foundation stones for the construction of covariance stationary models. The ARMA family of stationary models and other related processes that are discussed in this text possess covariance stationarity. If normality is assumed, second-order stationarity implies strict stationarity.

Because of the Wold decomposition theorem, stationary linear stochastic models possess the flexibility to model a wide range of natural time series. Nevertheless, as explained by authors such as Tong (1983) and Tong et al. (1985), nonlinear models can be useful in certain situations. In addition to linearity, models can also be classified according to properties of the theoretical ACF. Accordingly, both short (see Chapter 3) and long (refer to Part V) memory models are considered in the text and the relative usefulness of these classes of models is examined.

When employing a specified type of stochastic model to describe a natural time series, statistics other than the mean, variance and ACF may be important. For instance, when using a riverflow model for simulation studies in the design of a reservoir, statistics related to cumulative sums are important. This is because the storage in a reservoir is a function of the cumulative inflows less the outflows released by the dam. In particular, the importance of the rescaled adjusted range and Hurst coefficient in reservoir design, is discussed in Chapter 10. When considering situations where droughts or floods are prevalent, extreme value statistics should be entertained. Thus, practical engineering requirements necessitate the consideration of statistics that are directly related to the physical problem being studied.

PROBLEMS

2.1 In Section 2.2, a time series is defined. Based on your own experiences, write down three examples of continuous time series, equally spaced discrete time series, and unequally spaced discrete time series for which the variables being measured are continuous random variables.

2.2 A qualitative definition for a stochastic process is presented in Section 2.3. By referring to a book on stochastic processes, such as one of those referenced in Section 1.6.3, write down a formal mathematical definition for a stochastic process.

2.3 Stochastic processes are discussed in Section 2.3. Additionally, in Table 1.4.1 stochastic models are categorized according to the criteria of time (discrete and continuous) and state space (discrete and continuous). By utilizing books referenced in Sections 1.4.3 and 1.6.3, write down the names of three different kinds of stochastic models for each of the four classifications given in Table 1.4.1.

2.4 Strong and weak stationarity are discussed in Section 2.4.2. By referring to an appropriate book on stochastic processes, write down precise mathematical definitions for strong stationarity, weak stationarity of order k and covariance stationarity.

2.5 In Section 2.5, some basic statistical definitions are given. As a review of some other ideas for probability and statistics write down the definitions for a random variable, probability distribution function and cumulative distribution function. What is the exact definition for a Gaussian or normal probability distribution function? What is the central limit theorem and the weak law of large numbers? If you have forgotten some of the basic concepts in probability and statistics, you may wish to refer to an introductory text on probability and statistics to refresh your memory.

2.6 Ergodicity is briefly explained in Section 2.5.5. By referring to an appropriate book on stochastic processes or time series analysis, such as the one by Parzen (1962) or Hannan (1970), give a more detailed explanation of ergodicity than that presented in Section 2.5.5. Be sure that all variables used in any equations that you use in your presentation are clearly defined and explained.

2.7 Go to the library and take a look at the book by Wold (1954). Provide further details and insights about Wold's decomposition theorem which go beyond the explanation given in Section 2.7.

REFERENCES

The reader may also wish to refer to references on statistical water quality modelling, statistics, stochastic hydrology and time series analysis given at the end of Chapter 1.

DATA SETS

Environment Canada (1977). Historical streamflow summary, Ontario. Technical report, Water Survey of Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Canada.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysics Journal*, 126:152-158.

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society*, 79:242-261.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

HYDROLOGY

Ikeda, S. and Parker, G., Editors (1989). *River Meandering.* American Geophysical Union, Washington, D.C.

Klemes, V. (1974). The Hurst phenomenon: a puzzle? *Water Resources Research,* 10(4):675-688.

Speight, J. G. (1965). Meander spectra of the Angabunga River. *Journal of Hydrology,* 3:1-15.

STATISTICAL METHODS IN HYDROLOGY

Haan, C. T. (1977). *Statistical Methods in Hydrology.* The Iowa State University Press, Ames, Iowa.

McCuen, R. H. and Snyder, W. M. (1986). *Hydrologic Modeling: Statistical Methods and Applications.* Prentice-Hall, Englewood Cliffs, New Jersey.

Yevjevich, V. M. (1972a). *Probability and Statistics in Hydrology.* Water Resources Publications, Littleton, Colorado.

Yevjevich, V. M. (1972b). Structural analysis of hydrologic time series. Hydrology Paper No. 56, Colorado State University, Fort Collins, Colorado.

STATISTICS

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 26:211-252.

Guttman, I., Wilks, S. S., and Hunter, J. S. (1971). *Introductory Engineering Statistics.* John Wiley, New York, second edition.

Kalbfleisch, J. G. (1985). *Probability and Statistical Inference, Volume 1: Probability, Volume 2: Statistical Inference.* Springer-Verlag, New York.

Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis.* The Iowa State University Press, Ames, Iowa.

Kotz, S. and Johnson, N. L., Editors (1988). *Encyclopedia of Statistical Sciences.* Nine volumes, John Wiley, New York.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists.* John Wiley, New York.

Sachs, L. (1984). *Applied Statistics, A Handbook of Techniques.* Springer-Verlag, New York, second edition.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods.* Iowa State University Press, Ames, Iowa, seventh edition.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer sunspot numbers. *Phil. Transactions of the Royal Society, Series A,* 226:267-298.

STOCHASTIC PROCESSES

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes.* Chapman and Hall, London.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, second edition.

Parzen, E. (1962). *Stochastic Processes.* Holden-Day, San Francisco.

Ross, S. M. (1983). *Stochastic Processes.* John Wiley, New York.

TIME SERIES ANALYSIS

Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society, Series B,* 8:27-41.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

Cox, D. R. (1966). The null distribution of the first serial correlation coefficient. *Biometrika,* 53:623-626.

Cox, D. R. (1991). Long-range dependence, non-linearity and time irreversibility. *Journal of Time Series Analysis,* 12(4):329-335.

Hannan, E. J. (1970). *Multiple Time Series.* John Wiley, New York.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications.* Holden-Day, San Francisco.

McLeod, A. I. and Jimenez, C. (1984). Nonnegative definiteness of the sample autocovariance function. *The American Statistician,* 38(4):297.

Otnes, R. K. and Enochson, L. (1972). *Digital Time Series Analysis.* John Wiley, New York.

Robinson, E. A. (1967). *Multichannel Time Series Analysis with Digital Computer Programs.* Holden-Day, San Francisco.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Springer-Verlag, New York.

Tong, H., Thanoon, B., and Gudmundsson, G. (1985). Threshold time series modelling of two Icelandic riverflow systems. *Water Resources Bulletin,* 21(4):651-661.

Wold, H. (1954). *A Study in the Analysis of Stationary Time Series.* Almquist and Wicksell, Uppsala, Sweden, second edition.

PART II

LINEAR NONSEASONAL MODELS

Environmetrics is the development and application of statistical methodologies and techniques in the environmental sciences. As explained in Chapter 1 of Part I, statistical methods from the field of environmetrics can enhance scientific investigations of environmental problems and improve environmental decision making. Of primary interest in this book is the presentation of **useful time series models** that can be employed by water resources and environmental engineers for studying practical problems arising in hydrology and water quality modelling. Chapter 2 of Part I provides definitions and explanations for some important statistical techniques and concepts that are utilized in time series modelling and environmetrics.

The objectives of Part II of the book are to define a variety of **linear time series models** that can be applied to **nonseasonal time series** and to explain some of the key theoretical properties of these models which are required for understanding how to apply the models to actual data sets and to interpret the results. Chapters 3 and 4 describe linear nonseasonal models for fitting to **stationary and nonstationary time series**, respectively (see Section 2.4 for an explanation of stationarity and nonstationarity).

Figure II.1 displays a graph of the **annual flows** of the St. Lawrence River at Ogdensburg, New York, from 1860 to 1957. This figure is also given as Figure 2.3.1 in Chapter 2. The plotted time series appears to be **stationary** since statistical properties, such as the mean and variance, do not change over time. In addition, because there is no seasonal component, which would appear as some type of sinusoidal cycle in the graph, the data set is nonseasonal. The purpose of Chapter 3 is to describe three related families of linear time series models that could be considered for fitting to a time series like the one in Figure II.1. In particular, the three sets of models are:

1. **AR (autoregressive)** (Section 3.2),
2. **MA (moving average)** (Section 3.3), and
3. **ARMA (autoregressive-moving average) models** (Section 3.4).

The AR and MA models are in fact subsets of the general ARMA family of models. It turns out that the most appropriate model to fit to the yearly riverflow series of Figure II.1 is a special kind of AR model (Section 3.2.2). Indeed, within Section 3.6 it is demonstrated that there is sound **physical justifications** for fitting ARMA models to yearly riverflow time series.

The values of the **annual water usage** for New York City from 1898 to 1970 are plotted in Figure II.2 as well as Figure 4.3.8 in Chapter 4. Because the level of the series is increasing with time, the data are obviously **nonstationary**. Moreover, no seasonal cycle is contained in the graph. In Chapter 4, the following family of linear nonstationary time series models is described for applying to a nonstationary data set like the one in Figure II.2:

4. **ARIMA (autoregressive integrated moving average) models.**

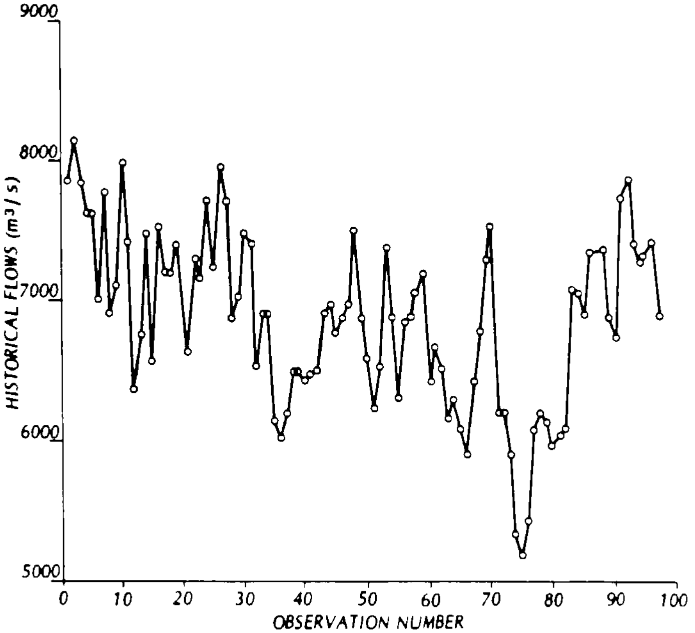


Figure II.1. Annual flows in m^3/s of the St. Lawrence River at Ogdensburg, New York, from 1860 to 1957.

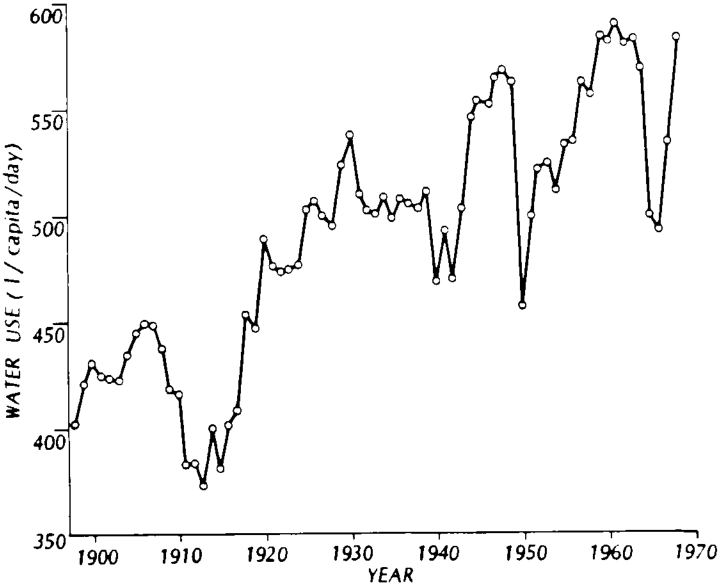


Figure II.2. Annual water use for New York City in litres per capita per day from 1898 to 1970.

When fitting an ARIMA model to a nonstationary series, the nonstationarity is removed from the series using a technique called **differencing**. Subsequently, appropriate AR and MA parameters contained in the ARIMA model are estimated for the resulting stationary series formed by differencing the original nonstationary series. In Section 4.3.3, it is explained how one can decide upon the most reasonable kind of ARIMA model to fit to the annual water use series for New York City.

The increasing levels of the water use series in Figure II.2 constitutes a trend in the data over time. **Deterministic and stochastic trends** are described in Section 4.6 along with approaches for modelling these types of trends. In fact, the ARIMA models of Chapter 4 constitute a procedure for modelling stochastic trends. The intervention models of Part VIII provide an approach for modelling known deterministic trends and estimating their magnitudes.

In summary, Part II of the book defines some **flexible families of linear nonseasonal models** for fitting to stationary (Chapter 3) and nonstationary (Chapter 4) time series. Additionally, useful theoretical properties for these models are pointed out so that a practitioner can decide upon or identify the most appropriate model to fit to a given time series. Part III describes how a user can fit the models of Part II to actual time series by following the identification (Chapter 5), estimation (Chapter 6), and diagnostic check (Chapter 7) stages of **model construction**. In fact, modified versions of the model building methods of Part III are employed with all of the kinds of time series models presented later in the book. Finally, techniques for **forecasting** and **simulating** using the models of Part II are given in Chapters 8 and 9, respectively, of Part IV.

CHAPTER 3

STATIONARY NONSEASONAL MODELS

3.1 INTRODUCTION

Certain types of environmental records are strictly *nonseasonal* while in other situations it may be required to consider a time series of average annual values even if seasonal data were available. For example, tree ring indices and mud varve thicknesses are usually obtainable only in the form of yearly records, whereas mean annual riverflow, temperature and precipitation data can be calculated from average weekly records. Whatever the case, it is often necessary to deal with nonseasonal natural time series.

The yearly data to be analyzed may be approximately stationary or perhaps may possess statistical properties which change over time. As discussed in Section 2.4.2, it is often reasonable to assume that hydrologic and geophysical data having a moderate time span (usually a few hundred years but perhaps more than 1000 years for certain time series) are more or less *stationary*. On the other hand, an annual water demand series for a large city or the yearly economic growth rate of an irrigated farming region, may constitute time series which are *nonstationary* even over a very short time interval. The present chapter deals with the theory of *stationary linear nonseasonal models* while Chapter 4 is concerned with *nonstationary linear nonseasonal models* which can be used for modelling certain types of nonstationary time series.

Nonseasonal models can be fit to yearly records for use in various types of applications. For instance, when *studying changes in the climate* over a specified time span, it may be advantageous to analyze annual time series. Although average annual hydrological data are rarely available for periods greater than two hundred years, longer time series, which reflect past climatic conditions, can be obtained. Some time series records of tree ring indices for the Bristlecone pine in California are longer than 5000 years in length and tree ring data sets for Douglas fir, Ponderosa pine, Jeffrey pine and other types of evergreens are available for periods of time which are often much longer than 500 years (Stokes et al., 1973).

Hurst (1951, 1956) studied the statistical properties of 690 annual time series when he was examining the long-term storage requirements on the Nile River. This research created the need for a stochastic model which could statistically account for what is called the *Hurst phenomenon*. Although the research of Hurst and accompanying academic controversies are assessed in detail in Chapter 10, it should be pointed out here that the linear stationary models of this chapter do in fact statistically explain the Hurst phenomenon (McLeod and Hipel, 1978; Hipel and McLeod, 1978). Consequently, stationary linear nonseasonal models are of great importance in hydrology and as emphasized in Chapter 10, should be employed in preference to fractional Gaussian noise (FGN) and other related models. Moreover, within Section 3.6 it is clearly demonstrated that there is sound *physical justification* for fitting the models of this chapter to yearly riverflow time series.

The current chapter deals with the *mathematical definitions and properties of various types of stationary linear nonseasonal processes*. The processes which are discussed are the *AR* (autoregressive), *MA* (moving average) and *ARMA* (autoregressive-moving average) processes.

For each of the foregoing processes, a simple process is first considered and this is followed by an extension to the general case. Important mathematical properties of the various processes are usually explained by examining a specific case. Furthermore, it is clearly pointed out where the mathematical properties of the processes can be useful for designing a model to fit to a given data set. The procedure of constructing a model by following the identification, estimation and diagnostic check system of *model development*, is discussed in Chapters 5 to 7, respectively, of Part III.

The importance of abiding by *key modelling principles* (see Sections 1.3 and 5.2.4 for general discussions) is addressed at certain locations within this chapter. For example, in order to make the model as simple or parsimonious as possible, some of the model parameters can be constrained to zero (see Section 3.4.4). To satisfy certain underlying modelling assumptions regarding the model residuals, a power transformation such as a *Box-Cox transformation* (Box and Cox, 1964) can be incorporated into the model (see Section 3.4.5).

The mathematical foundations of linear nonseasonal models form the *basic building blocks* for the more complex nonstationary, long memory, seasonal, transfer function-noise, intervention and multivariate models which are dealt with in Chapters 4 and 11, and Parts VI to IX, respectively, later in the book. Consequently, a sound understanding of the models presented in this chapter is essential in order to be able to fully appreciate the flexibility and limitations of the rich array of ARMA-based models which are available for use by engineers. In addition, the basic notation which is developed for the nonseasonal models is simply extended for use with the other classes of models described in the book.

3.2 AUTOREGRESSIVE PROCESSES

The AR model of this section describes how an observation directly depends upon one or more previous measurements plus a white noise term. This form of a time series model is intuitively appealing and has been widely applied to data sets in many different fields. After describing the simplest form of the AR model, the general AR model is defined. Additionally, the theoretical ACF (autocorrelation function) of an AR model is derived and the related Yule-Walker equations are formulated. These equations can be used for obtaining the partial autocorrelation function (PACF) and determining efficient moment estimates for the parameters of an AR model.

3.2.1 Markov Process

When an observation, z_t, measured at time t depends only upon the time series value at time t-1 plus a random shock, a_t, the process describing this relationship is called an AR process of order 1 and is denoted as AR(1). The AR(1) process is commonly called a *Markov process* and is written mathematically as

$$z_t - \mu = \phi_1(z_{t-1} - \mu) + a_t \qquad [3.2.1]$$

where μ is the mean level of the process, ϕ_1 is the nonseasonal AR parameter, a_t is the white noise term at time t that is *identically independently distributed (IID)* with a mean of 0 and variance of σ_a^2 [i.e. IID $(0,\sigma_a^2)$].

The a_t sequence is referred to as *random shocks, disturbances, innovations* or *white noise terms*. After a model has been fit to a given time series and estimates have been obtained for the innovations, the estimates are called estimated innovations or *residuals*.

The most important assumption for the random shocks is that they are independently distributed. This infers that the a_t's are uncorrelated and must satisfy

$$E[a_t a_{t-k}] = \begin{cases} \sigma_a^2, & k = 0 \\ 0, & k \neq 0 \end{cases} \qquad [3.2.2]$$

The a_t's follow the same distribution and sometimes it is convenient to assume that the random shocks are normally distributed. This may be appropriate for estimation purposes, forecasting and simulation. In addition, if normal random variables are uncorrelated then they are also independent.

The difference equation in [3.2.1] can be written more economically by introducing the *backward shift operator B* which is defined by

$$Bz_t = z_{t-1}$$

and

$$B^k z_t = z_{t-k}$$

where k is a positive integer. By using the B operator, the Markov process in [3.2.1] is

$$z_t - \mu = \phi_1(Bz_t - \mu) + a_t$$

or

$$z_t - \mu - \phi_1(Bz_t - \mu) = a_t$$

By treating B as an algebraic operator and factoring, the above equation becomes

$$(1 - \phi_1 B)(z_t - \mu) = a_t$$

where $B\mu = \mu$ since the mean level is a constant at all times. The previous equation can also be given as

$$\phi(B)(z_t - \mu) = a_t \qquad [3.2.3]$$

where $\phi(B) = 1 - \phi_1 B$ is the nonseasonal AR operator or polynomial of order one.

3.2.2 Autoregressive Process of Order p

The Markov process with the single AR parameter, ϕ_1, is a special case of an *AR process of order p [i.e. AR(p)]* which is given as

$$z_t - \mu = \phi_1(z_{t-1} - \mu) + \phi_2(z_{t-2} - \mu) + \cdots + \phi_p(z_{t-p} - \mu) + a_t \qquad [3.2.4]$$

where ϕ_i is the ith *nonseasonal AR parameter*. By introducing the B operator, [3.2.4] can equivalently be written as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(z_t - \mu) = a_t$$

or

$$\phi(B)(z_t - \mu) = a_t \qquad [3.2.5]$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the *nonseasonal AR operator* of order p.

Stationarity

The equation $\phi(B) = 0$ is referred to as the *characteristic equation* for the process. It can be shown (Box and Jenkins, 1976, Ch. 3; Pagano, 1973) that a necessary and sufficient condition for the process to have *stationarity* is that the roots of the characteristic equation must fall outside the *unit circle*. The unit circle is a circle of unit radius centered on the origin of a complex number graph where one axis is the real number component and the other axis forms the imaginary part of the complex number.

Based upon the work of Schur (1917), Pagano (1973) presented an algorithm which can be used to determine whether or not all the roots of a given operator lie outside the unit circle. Consider the situation where it is necessary to ascertain if all the roots of the operator, $\phi(B)$, for an AR(p) process fall outside the unit circle. The first step is to form the Schur matrix **A** of dimension p×p which has (i,j)th element

$$\sum_{k=0}^{min(i,j)} (\phi_{i-k-1}\phi_{j-k-1} - \phi_{p+1+k-i}\phi_{p+1+k-j})$$

where $\phi_{-k} = 0$ and $\phi_0 = 1$. The matrix **A** is actually the inverse of the covariance matrix of p successive observations for an AR(p) process (Siddiqui, 1958). Schur (1917) demonstrated that a necessary and sufficient condition for the roots of $\phi(B) = 0$ to lie outside the unit circle is for **A** to be positive definite. Because **A** is positive definite whenever the covariance matrix is positive definite (Pagano, 1973), to demonstrate that an AR(p) process is stationary it is only necessary to show that **A** is positive definite. A convenient way to do this is to calculate the *Cholesky decomposition* of **A** [see Wilkinson (1965) and Healy (1968)] given by

$$\mathbf{A} = \mathbf{M}\,\mathbf{M}^T$$

where **M** is a lower triangular matrix. If all the diagonal entries of **M** are positive, matrix **A** is positive definite. When there are one or more zero entries on the diagonal of **M** and all other entries are positive, **A** is positive semidefinite. If during the calculation of **M** a diagonal location is encountered where a zero or positive entry cannot be calculated, the Cholesky decomposition does not exist. However, when the Cholesky decomposition shows that **A** is positive definite, then the roots of $\phi(B) = 0$ lie outside the unit circle. For the case of the $\phi(B)$ operator, this property means that the process is stationary.

From [3.2.3] the characteristic equation for the Markov process is

$$(1 - \phi_1 B) = 0$$

By considering B as an algebraic variable, the root of the characteristic equation is $B = \phi_1^{-1}$. In order for ϕ_1^{-1} to lie outside the unit circle to ensure stationarity, then $|\phi_1| < 1$.

The stationarity condition automatically ensures that a process can be written in terms of the a_t's in what is called a pure MA process. For example, the AR(1) process in [3.2.3] can be expressed as

$$z_t - \mu = (1 - \phi_1 B)^{-1} a_t \qquad [3.2.6]$$
$$= (1 + \phi_1 B + \phi_1^2 B^2 + \cdots) a_t$$

Because $|\phi_1| < 1$ due to the stationarity condition, this infers that the infinite series $(1 - \phi_1 B)^{-1}$ will converge for $|B| \leq 1$. The beneficial consequences caused by the restriction upon ϕ_1 can also be explained by writing [3.2.6] as

$$z_t - \mu = a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \phi_1^3 a_{t-3} + \cdots$$

If $|\phi_1| < 1$, the dependence of the deviation $(z_t - \mu)$ upon the white noise terms decreases further into the past. Alternatively, if $|\phi_1| \geq 1$, the dependence of $(z_t - \mu)$ upon the white noise would be greater for disturbances which happened well before the more recent shocks. Of course, this type of interpretation would not be meaningful for stationary processes and can be avoided if the stationarity condition is satisfied.

Autocorrelation Function

In order to study the properties of the theoretical ACF for a stationary AR(p) process, firstly multiply [3.2.4] by $(z_{t-k} - \mu)$ to obtain

$$(z_{t-k} - \mu)(z_t - \mu) = \phi_1(z_{t-k} - \mu)(z_{t-1} - \mu) + \phi_2(z_{t-k} - \mu)(z_{t-2} - \mu) + \cdots$$
$$+ \phi_p(z_{t-k} - \mu)(z_{t-p} - \mu) + (z_{t-k} - \mu)a_t \qquad [3.2.7]$$

By taking expected values of [3.2.7], the difference equation for the *autocovariance function of the AR(p) process* is

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p}, \quad k > 0 \qquad [3.2.8]$$

The term $E[(z_{t-k} - \mu)a_t]$ is zero for $k > 0$ because z_{t-k} is only a function of the disturbances up to time $t-k$ and a_t is uncorrelated with these shocks. To determine an expression for the theoretical *ACF for the AR(p) process*, divide [3.2.8] by γ_0 to obtain

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p}, \quad k > 0$$

This equation can be equivalently written as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\rho_k = \phi(B)\rho_k = 0, \quad k > 0 \qquad [3.2.9]$$

where B operates on k instead of t. The general solution of the difference equation in [3.2.9] is (Box and Jenkins, 1976, p. 55)

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad [3.2.10]$$

where $G_1^{-1}, G_2^{-1}, \cdots, G_p^{-1}$, are distinct roots of the characteristic equation $\phi(B) = 0$ and the A_i's are constants. If a root G_i^{-1} is real then $|G_i^{-1}|>1$ due to the stationarity conditions. Hence,

$|G_i|<1$ and $A_i G_i^k$ in [3.2.10] forms a damped exponential which geometrically decays to zero as k increases. Complex roots contribute a damped sine wave to the theoretical ACF in [3.2.10]. Consequently, the theoretical ACF for a stationary AR process will consist of a combination of damped exponential and sine waves.

St. Lawrence River Data: As mentioned in Section 2.5.4, when determining a model to fit to a given data set, it is desirable to have the theoretical ACF of the process to resemble statistically the sample ACF. Consider, for example, the average annual flows of the St. Lawrence River at Ogdensburg, New York. These flows are available from 1860 to 1957 in a report by Yevjevich (1963). The estimated ACF for these yearly flows is calculated using [2.5.9] and is shown in Figure 3.2.1. The 95% confidence limits are determined utilizing [2.5.11] by assuming that the sample ACF is not significantly different from zero after lag 0. As can be seen in Figure 3.2.1, the estimated ACF has significant non-zero values at lower lags and tends to follow a damped exponential curve. Because the theoretical ACF of an AR process behaves in this fashion, this indicates that perhaps some type of model which contains an AR component should be fit to the St. Lawrence flows.

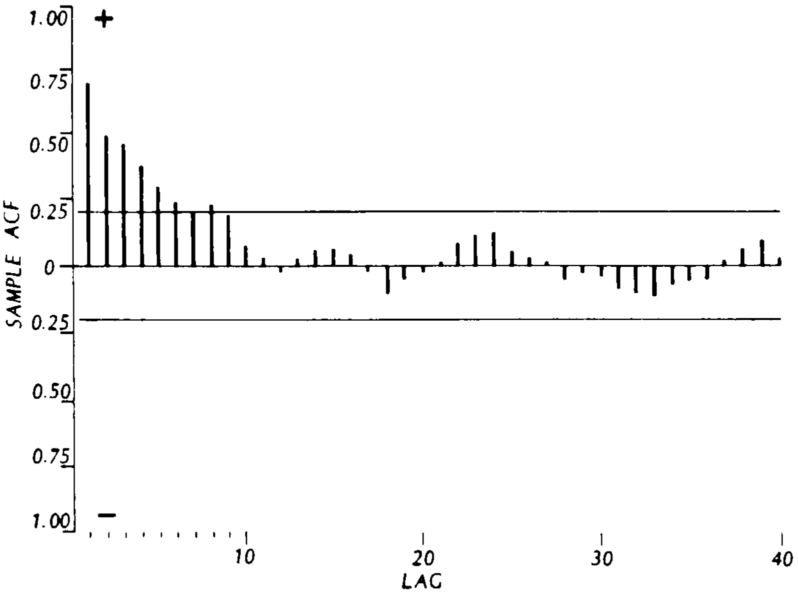


Figure 3.2.1. Sample ACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

Yule-Walker Equations

By substituting $k = 1,2, \ldots, p$, into [3.2.9], parameters can be expressed in terms of the theoretical ACF. The resulting set of linear equations are called the *Yule-Walker equations* [after Yule (1927) and Walker (1931)] and are given by

$$\begin{array}{l} \rho_1 = \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1} \\ \rho_2 = \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2} \\ \quad . \quad . \quad . \quad \cdots \quad . \\ \quad . \quad . \quad . \quad \cdots \quad . \\ \quad . \quad . \quad . \quad \cdots \quad . \\ \rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p \end{array} \qquad [3.2.11]$$

By writing the Yule-Walker equations in matrix form, the relationship for the AR parameters is

$$\phi = \mathbf{P}_p^{-1}\rho_p \qquad [3.2.12]$$

where

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ . \\ . \\ . \\ \phi_p \end{bmatrix}, \quad \rho_p = \begin{bmatrix} \rho_1 \\ \rho_2 \\ . \\ . \\ . \\ \rho_p \end{bmatrix}, \quad \mathbf{P}_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}$$

To obtain Yule-Walker estimates for the AR parameters, simply replace the ρ_k's in [3.2.12] by their estimates r_k, $k = 1,2,\ldots,p$, from [2.5.9]. The Yule-Walker estimates possess large sample efficiency and hence have minimum possible variances.

By setting $k = 0$ in [3.2.7] and taking expectations, the expression for the variance is

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma_a^2 \qquad [3.2.13]$$

where $E[z_t a_t] = \sigma_a^2$ since z_t is only correlated with a_t due to the most recent shock a_t. Upon dividing [3.2.13] by $\gamma_0 = \sigma_z^2$, the variance of the process can be expressed as

$$\sigma_z^2 = \frac{\sigma_a^2}{1 - \rho_1\phi_1 - \rho_2\phi_2 - \cdots - \rho_p\phi_p} \qquad [3.2.14]$$

Employing [3.2.13] and [2.5.8], the residual variance can be estimated using

$$\hat{\sigma}_a^2 = c_o - \sum_{i=1}^{p}\hat{\phi}_i c_i$$

In addition to the Yule-Walker estimator, other estimators are available for efficiently estimating the parameters of an AR model. One approach is to employ the maximum likelihood estimator presented in Section 6.2 and Appendix A6.1. A second procedure is to employ the Burg (1975) algorithm which is described by Haykin (1990, pp. 187-192).

Markov Process: As shown earlier in this section, in order for an AR(1) process to be stationary $-1<\phi_1<1$. By setting ϕ_2 to ϕ_p equal to zero, equation [3.2.11] becomes

$$\rho_1 = \phi_1$$

$$\rho_2 = \phi_1 \rho_1 = \phi_1^2$$

$$\rho_3 = \phi_1 \rho_2 = \phi_1^3$$

In general,

$$\rho_k = \phi_1^k. \qquad [3.2.15]$$

Because of the form of [3.2.15], the theoretical ACF attenuates exponentially to zero if ϕ_1 is positive but decays exponentially to zero and oscillates in sign when ϕ_1 is negative. From Figure 3.2.2, it can be seen that when ϕ_1 is assigned a positive value of 0.75, the theoretical ACF only possesses positive values which decay exponentially to zero for increasing lag. However, when ϕ_1 is given a negative value such as -0.75, the theoretical ACF oscillates in sign and decays exponentially to zero as shown in Figure 3.2.3. The variance of an AR(1) process is obtained from [3.2.14] and [3.2.15] as

$$\sigma_z^2 = \frac{\sigma_a^2}{1 - \rho_1 \phi_1} = \frac{\sigma_a^2}{1 - \phi_1^2} \qquad [3.2.16]$$

Partial Autocorrelation Function

Because the ACF of an AR process attenuates and does not truncate at a specified lag, it would be advantageous to define a function which does cut off for an AR process. As explained in Chapter 5, such a device would be useful to employ in conjunction with the sample ACF and other tools for identifying the type of model to fit to a given data set.

Let ϕ_{kj} be the jth coefficient in a stationary AR process of order k so that ϕ_{kk} is the last coefficient. The Yule-Walker equations in [3.2.12] can then be equivalently written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ . \\ . \\ . \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ . \\ . \\ . \\ \rho_k \end{bmatrix} \qquad [3.2.17]$$

The coefficient ϕ_{kk} is a function of the lag k and is called the theoretical *partial autocorrelation function (PACF)*. Because of the definition of the theoretical PACF, it must be equal to zero after lag p for an AR(p) process. Furthermore, the possible values of ϕ_{kk} range from -1 to 1.

One method for estimating the PACF is to employ the Yule-Walker equations. By replacing ρ_k in [3.2.17] by its estimate r_k from [2.5.9], the estimates of ϕ_{kk}, $k = 1,2,...,$ using Cramer's rule are

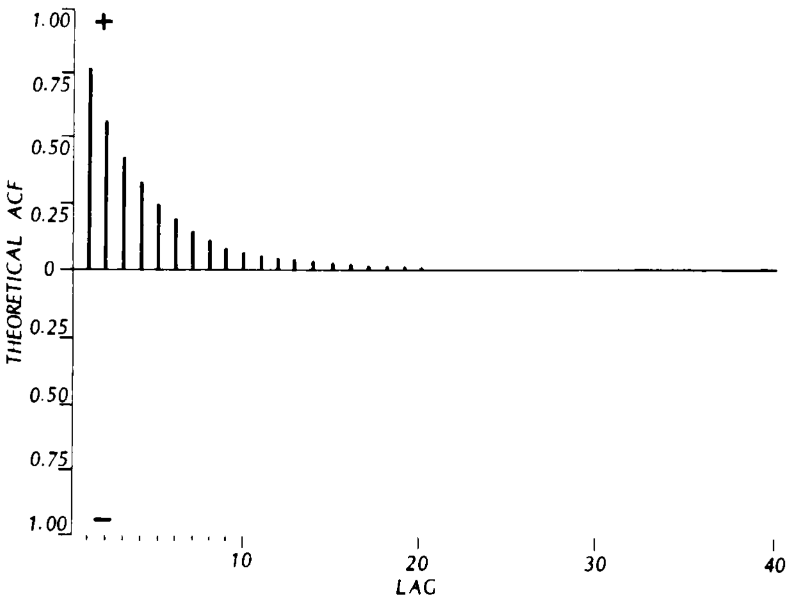


Figure 3.2.2. Theoretical ACF for a Markov Process with $\phi_1 = 0.75$.

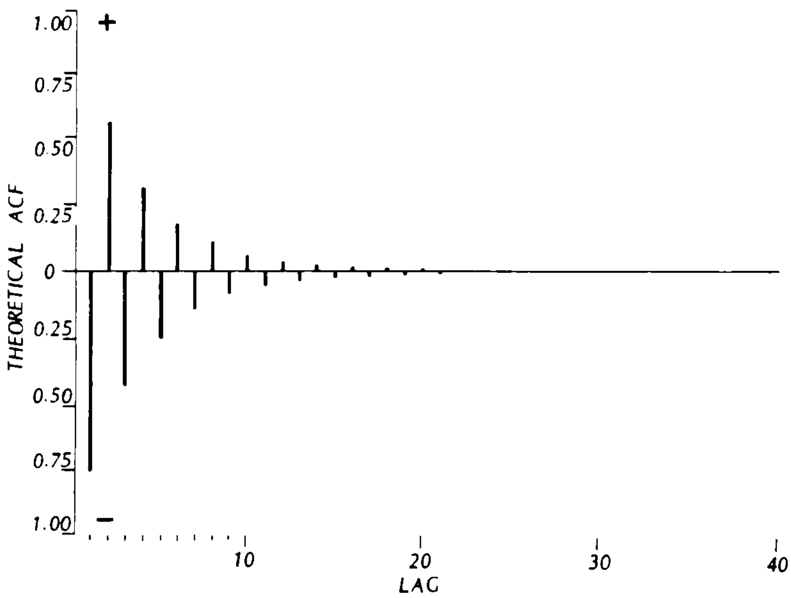


Figure 3.2.3. Theoretical ACF for a Markov Process with $\phi_1 = -0.75$.

$$\hat{\phi}_{11} = r_1,$$

$$\hat{\phi}_{22} = \left[\frac{\begin{matrix}1 & r_1\\ r_1 & r_2\end{matrix}}{\begin{matrix}1 & r_1\\ r_1 & 1\end{matrix}}\right] = \frac{r_2-r_1^2}{1-r_1^2}$$

etc. In order to make the estimation of the PACF computationally more economical, the recursive formulae of Durbin (1960) may be employed. However, as noted by Pagano (1972) and Box and Jenkins (1976), Durbin's method is numerically unstable, especially when the process approaches nonstationarity (i.e., the roots of characteristic equation are close to the unit circle).

An alternative procedure for estimating the PACF is to utilize the algorithm devised by Pagano (1972). The *Pagano algorithm* is numerically quite stable because it is based upon the Cholesky decomposition which is known to be stable (Wilkinson, 1965, pp. 231 and 244). The steps required in the Pagano algorithm for calculating the PACF up to lag p are given in Appendix A3.1. Because the algorithm is numerically stable and is also economical with respect to computational requirements, it is amenable for programming on the computer.

When plotting $\hat{\phi}_{kk}$ against lag k, approximate confidence limits must be given in order to decipher values of the estimated PACF which are significantly different from zero. If the process is AR(p), the sample PACF should not be significantly different from zero after lag p. Based upon the hypothesis that the process is AR(p), the estimated values of the PACF at lags greater than p are approximately normally independently distributed with a SE given by (Quenouille, 1949; Barndorff-Nielsen and Schou, 1973)

$$\mathrm{SE}[\hat{\phi}_{kk}] \approx \frac{1}{\sqrt{N}} \qquad [3.2.18]$$

where N is the length of the time series.

St. Lawrence River Data: The graph of the estimated PACF for the average annual flows of the St. Lawrence River is shown in Figure 3.2.4. The 95% confidence limits are calculated by substituting $N = 97$ into [3.2.18] and plotting 1.96 times the SE for $\hat{\phi}_{kk}$ above and below the horizontal axis. It can be seen that there are rather large values for the estimated PACF at lags 1, 3 and 19. The unexpected big value at lag 19 could be due to chance alone or else the limited size of the sample which was used to estimate the PACF at lag 19. Because the estimated PACF cuts off after lag 3, this implies that an AR(3) process should perhaps be fitted to the data. In addition, because the sample PACF at lag 2 is not very large, perhaps the ϕ_2 parameter should be constrained to zero in the AR(3) model in order to reduce the number of model parameters. As shown in Section 6.4.2, the estimated model for the St. Lawrence data is

$$(1 - 0.619B - 0.177B^3)(z_t - 6818.63) = a_t \qquad [3.2.19]$$

where 6818.63 is the maximum likelihood estimate for the mean.

By substituting the values of the AR parameters for the model from [3.2.19] into the Yule-Walker equations in [3.2.11], the theoretical ACF can be determined. It can be seen from Figure 3.2.5 that the theoretical ACF for the St. Lawrence model in [3.2.19] is statistically similar to the

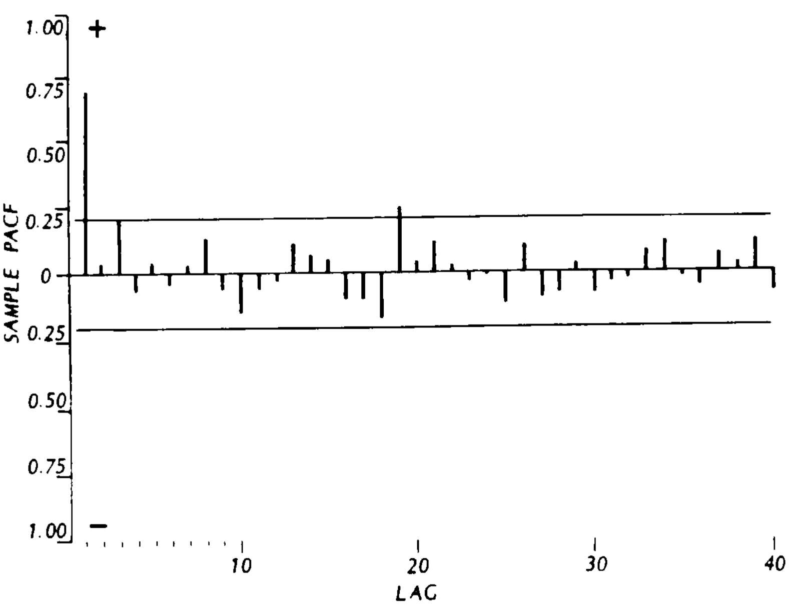


Figure 3.2.4. Sample PACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

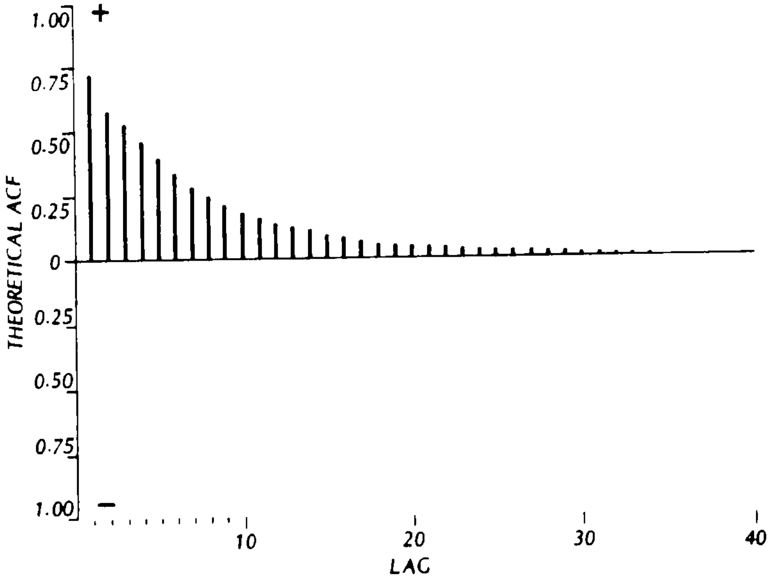


Figure 3.2.5. Theoretical ACF for the AR(3) model without ϕ_2 that is fitted to the average annual flows of the St. Lawrence River at Ogdensburg, New York.

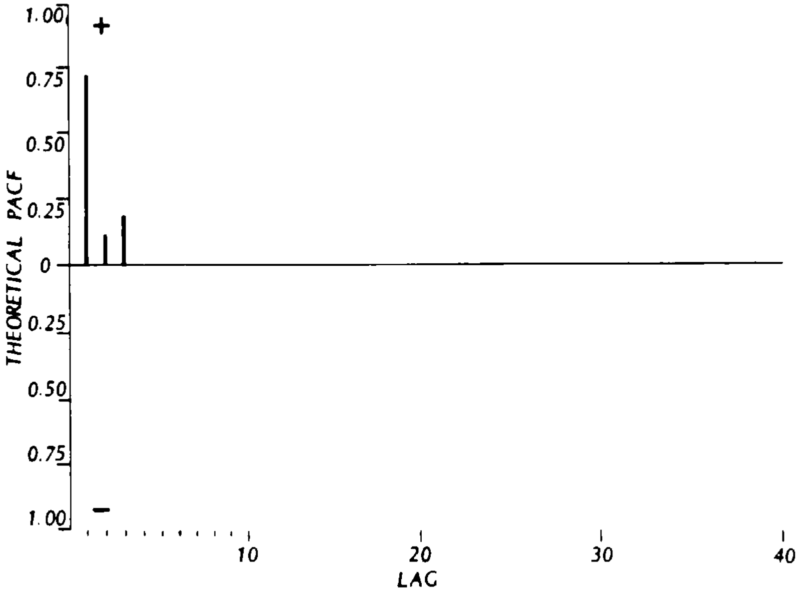


Figure 3.2.6. Theoretical PACF for the AR(3) model without ϕ_2 that is fitted to the average annual flows of the St. Lawrence River at Ogdensburg, New York.

sample ACF given in Figure 3.2.1. This information indicates that an AR model is a reasonable type of model to fit to the St. Lawrence River flows.

To further justify the use of the model in [3.2.19] for modelling the St. Lawrence River flows, the theoretical PACF can be compared to the sample PACF in Figure 3.2.4. In order to calculate the theoretical PACF, the values of theoretical ACF which were determined by substituting the estimates for the AR parameters in [3.2.19] into [3.2.11], are employed in [3.2.17]. The graph of the theoretical PACF for the St. Lawrence River flows is shown in Figure 3.2.6 and it can be seen that this plot is similar to the sample PACF in Figure 3.2.4.

3.3 MOVING AVERAGE PROCESSES

The MA model describes how an observation depends upon the current white noise term as well as one or more previous innovations. After examining the simplest type of MA model, the general form of the MA model is defined and its important theoretical properties are derived.

3.3.1 First Order Moving Average Process

When a time series value, z_t, is dependent only upon the white noise at time $t-1$ plus the current shock, the relationship is written as

$$z_t - \mu = a_t - \theta_1 a_{t-1} \qquad [3.3.1]$$

where θ_1 is the nonseasonal MA parameter. This process is termed a MA process of order one and is denoted as MA(1). By introducing the B operator, the MA(1) process can be equivalently written as

$$\begin{aligned} z_t - \mu &= a_t - \theta_1 B a_t \\ &= (1 - \theta_1 B) a_t \\ &= \theta(B) a_t \end{aligned} \qquad [3.3.2]$$

where $\theta(B) = 1 - \theta_1 B$ is the nonseasonal MA operator or polynomial of order one.

3.3.2 Moving Average Process of Order q

The MA(1) process can be readily extended to the situation where there are q MA parameters. The *MA process* of order q is denoted by MA(q) and is written as

$$z_t - \mu = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad [3.3.3]$$

where θ_j is the jth nonseasonal MA parameter. By employing the B operator, the MA(q) process can be more economically presented as

$$\begin{aligned} z_t - \mu &= a_t - \theta_1 B a_t - \theta_2 B^2 a_t - \cdots - \theta_q B^q a_t \\ &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t \\ &= \theta(B) a_t \end{aligned} \qquad [3.3.4]$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, is the *nonseasonal MA operator* or polynomial of order q.

Stationarity

The time series composed of the a_t's is assumed to be stationary. Because z_t in [3.3.4] is formed by a finite linear combination of the a_t's, then z_t must be stationary no matter what values the MA parameters possess. However, it is advantageous to put certain restrictions upon the range of values for the MA parameters. Consider, for example, the MA(1) process in [3.3.2]. By invoking the binomial theorem, this process can be equivalently written as an infinite AR process given as

$$\begin{aligned} a_t &= (1 - \theta_1 B)^{-1} (z_t - \mu) \\ &= (1 + \theta_1 B + \theta_1^2 B^2 + \theta_1^3 B^3 + \cdots)(z_t - \mu) \end{aligned} \qquad [3.3.5]$$

In order for the infinite series $(1 - \theta_1 B)^{-1}$ to converge for $|B| \le 1$, the parameter θ_1 must be restricted to have an absolute value less than unity. Another way to interpret the restriction upon θ_1 is to write [3.3.5] as

$$z_t - \mu = a_t - \theta_1 (z_{t-1} - \mu) - \theta_1^2 (z_{t-2} - \mu) - \theta_1^3 (z_{t-3} - \mu) - \cdots \qquad [3.3.6]$$

If $|\theta_1| > 1$, it can be seen in [3.3.6] that the current deviation $(z_t - \mu)$ depends more on events that happened further in the past because θ_1^k increases as the lag k gets larger. When $|\theta_1| = 1$, something that took place a long time ago has as much influence as a recent observation upon the current measurement. In order to avoid these situations, it is necessary that $|\theta_1| < 1$. This is equivalent to stipulating that the root $B = \theta_1^{-1}$ of the characteristic equation $(1 - \theta_1 B) = 0$ must lie outside the unit circle. Consequently, the stationary MA(1) process can only be meaningfully expressed as an infinite AR process if a restriction is placed upon the MA parameter. This restriction is referred to as the invertibility condition and is independent of the stationarity requirements of a process.

Invertibility

The characteristic equation for a MA(q) process is

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0 \qquad [3.3.7]$$

In order for a MA(q) process to be *invertible,* the roots of the characteristic equation must lie outside the unit circle.

An inherent advantage of the invertibility condition is that it does not cause a loss in generality of the MA process. As shown by Fuller (1976, pp. 64-66) and discussed by Anderson (1971, p. 204), any finite MA process whose characteristic equation has some roots greater than one and some less than one can be given a representation whose characteristic equation has all roots greater than one in absolute value. Consequently, the invertibility condition does not limit the ability to identify a suitable invertible model to fit to a given series. In addition, if the invertibility condition is satisfied, a MA process can be expressed as a pure AR process. Finally, when the residuals are being established for a model which is being fitted to a specified time series, the calculation of the residuals will be ill-conditioned if the invertibility condition is not met.

Autocorrelation Function

By using [2.5.3] and [3.3.3], the *autocovariance function of a MA(q) process* is

$$\begin{aligned}\gamma_k &= E[(z_t - \mu)(z_{t-k} - \mu)]\\ &= E[(a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q})(a_{t-k} - \theta_1 a_{t-k-1} - \theta_2 a_{t-k-2}\\ &\qquad - \cdots - \theta_q a_{t-k-q})]\end{aligned} \qquad [3.3.8]$$

After multiplication and taking expected values, the autocovariance function is

$$\gamma_k = \begin{cases} (-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q)\sigma_a^2 \quad , \quad k = 1,2,\ldots,q \\ \\ 0 \quad , \quad k>q \end{cases} \qquad [3.3.9]$$

where $\theta_0 = 1$ and $\theta_{-k} = 0$ for $k \geq 1$. When k is set equal to zero in [3.3.8], the variance is

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_a^2 \qquad [3.3.10]$$

By dividing the autocovariance function by the variance, the theoretical *ACF for a MA(q) process* is found to be

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} , & k = 1,2,\ldots,q \\ 0 , & k > q \end{cases} \qquad [3.3.11]$$

Partial Autocorrelation Function

It is shown in [3.3.5] that a MA(1) process can be equivalently written as an infinite AR process. In general, any finite invertible MA process can be expressed as an infinite AR process. Because the PACF is theoretically defined to be zero after lag p for a finite AR(p) process, the PACF must therefore attenuate at increasing lags for a MA process or equivalently an infinite AR process.

Temperature Data: From [3.3.11], it can be seen that the theoretical ACF for a MA(q) process is exactly zero after lag q. If the sample ACF is tabulated for a given time series using [2.5.9], then the estimated ACF should not be significantly different from zero after lag q if the underlying process is MA(q). For instance, the sample ACF for the average annual temperature data from the English Midlands is shown in Figures 2.5.1 and 2.5.2 in Section 2.5.4. Because the estimated ACF is not significantly different from zero after lag 2, it is reasonable to fit a MA(2) model to the data. Using the estimator described in Appendix A6.1, the estimated model for the temperature data is found to be

$$z_t - 9.216 = (1 + 0.111B + 0.197B^2)a_t \qquad [3.3.12]$$

By substituting the estimates for the MA parameters from [3.3.12] into [3.3.11] (where $\rho_k = 0$ for $k > 2$), the theoretical ACF can be calculated for the MA(2) model. By comparing the theoretical ACF in Figure 2.5.3 to the estimated ACF in Figure 2.5.1, it can be seen that the theoretical ACF for the MA(2) model mimics the estimated ACF.

To calculate the sample PACF for the temperature data from the English Midlands, one first must determine the sample ACF using [2.5.9]. Following this, Pagano's (1972) algorithm, outlined in Appendix A3.1, can be used to solve [3.2.17] in order to determine the sample PACF. The sample PACF along with the 95% confidence limits for the English temperature data, are displayed in Figure 3.3.1. The sample PACF truncates after lag 2 except for a rather large value at lag 15 which is probably due to chance. However, the plot of the sample ACF in Figure 2.5.2 reveals that it also cuts off after lag 2. Hence, either a MA(2) or an AR(2) model may adequately model the temperature data. As is shown in Section 3.4.3 the two models are in fact shown to be almost the same by expressing the AR(2) model as an infinite MA model in which the coefficients after lag 2 are negligible.

After substituting the values of the theoretical ACF for the MA(2) model in [3.3.12] into [3.2.17], one can employ Pagano's (1972) algorithm outlined in Appendix A3.1 to determine the theoretical PACF. It can be seen that the theoretical PACF in Figure 3.3.2 for the estimated MA(2) model closely resembles the sample PACF of the temperature data in Figure 3.3.1.

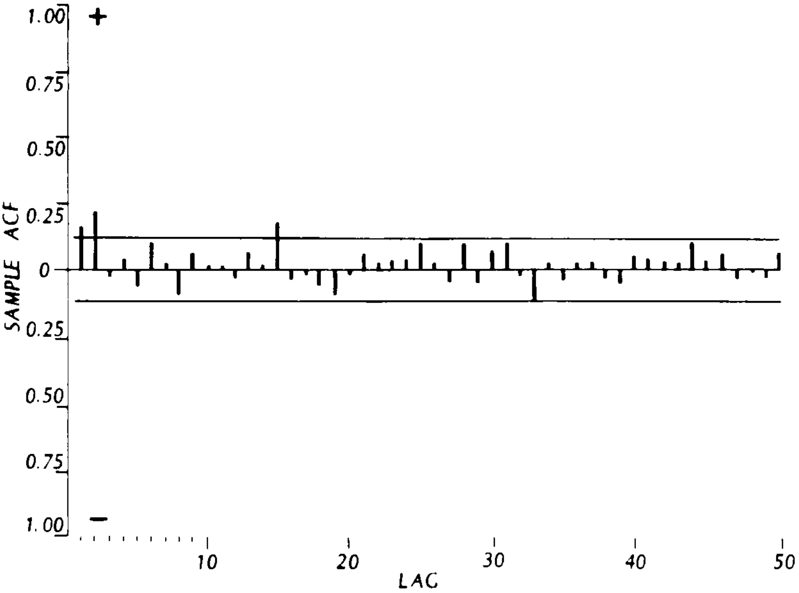


Figure 3.3.1. Sample PACF and 95% confidence limits for the annual temperature data from the English Midlands.

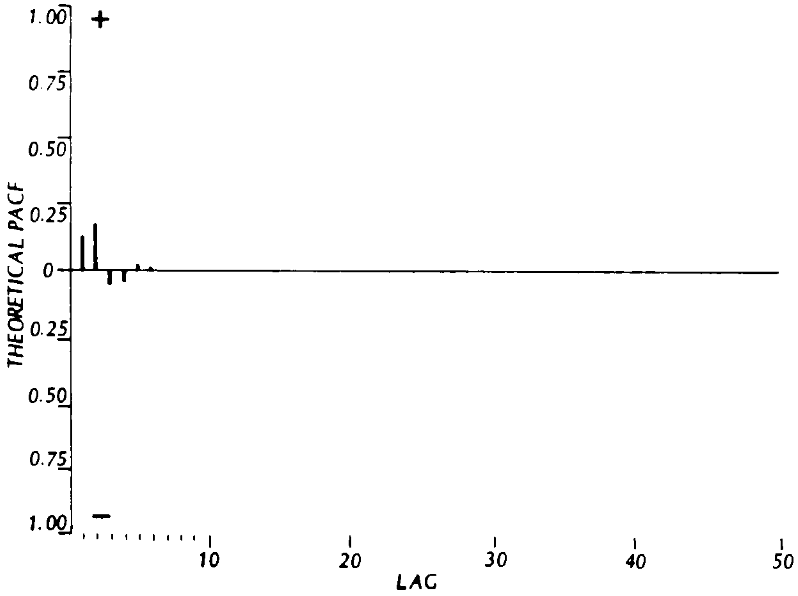


Figure 3.3.2. Theoretical PACF for the MA(2) model fitted to the annual temperature data from the English Midlands.

First Order Moving Average Process

The MA(1) process is given in [3.3.2]. This process is stationary for all values of θ_1 but for invertibility $|\theta_1| < 1$. When the invertibility condition is satisfied, the MA(1) process can be equivalently written as an infinite AR process as is shown in [3.3.5].

By utilizing [3.3.10], the variance of the MA(1) process is

$$\gamma_0 = (1 + \theta_1^2)\sigma_a^2 \tag{3.3.13}$$

From [3.3.11] the theoretical ACF is

$$\rho_k = \begin{cases} \dfrac{-\theta_1}{1+\theta_1^2} \ , & k = 1 \\ \\ 0 \ , & k \geq 2 \end{cases} \tag{3.3.14}$$

By substituting $\rho_1 = -\theta_1/(1 + \theta_1^2)$ and $\rho_k = 0$, for $k > 1$, into [3.2.17], the PACF can be shown to be

$$\phi_{kk} = -\theta_1^k(1 - \theta_1^2)/(1 - \theta_1^{2(k+1)}) \tag{3.3.15}$$

Because $|\theta_1| < 1$ for an invertible MA(1) process, the theoretical PACF decreases in value for increasing lag and follows a damped exponential curve. Because of the form of [3.3.15], it can be seen that $|\phi_{kk}| < |\theta_1|^k$. When θ_1 is positive, then from [3.3.14] ρ_1 is negative and the PACF values in [3.3.15] are also negative. On the other hand, if θ_1 is negative, ρ_1 is positive and the PACF values alternate in sign.

3.4 AUTOREGRESSIVE-MOVING AVERAGE PROCESSES

As noted in Sections 1.3 and 5.2.4, a key modelling principle is to have as few parameters as possible in the model. If, for example, the sample ACF for a given data set possesses a value which is significantly different from zero only at lag one, then it may be appropriate to fit a MA(1) model to the data. An AR model may require quite a few AR parameters in order to adequately model the same time series. When the sample PACF for another data set cuts off at lag 2, then the most parsimonious model to fit the time series may be an AR(2) model. In situations where both the sample ACF and PACF attenuate for a certain time series, it may be advantageous to have a model which contains both AR and MA parameters. In this way, the fitted model can be kept as simple as possible by keeping the number of model parameters to a minimum.

3.4.1 First Order Autoregressive-First Order Moving Average Process

If a process consists of both AR and MA parameters, it is called an ARMA process. When there is one AR and one MA parameter the ARMA process is denoted as ARMA(1,1) and the equation for this process is

$$(z_t - \mu) - \phi_1(z_{t-1} - \mu) = a_t - \theta_1 a_{t-1} \qquad [3.4.1]$$

By utilizing the B operator, the ARMA(1,1) process can be equivalently written as

$$(1 - \phi_1 B)(z_t - \mu) = (1 - \theta_1 B)a_t$$

or

$$\phi(B)(z_t - \mu) = \theta(B)a_t \qquad [3.4.2]$$

where $\phi(B) = 1 - \phi_1 B$ and $\theta(B) = 1 - \theta_1 B$ are, respectively, the AR and MA operators of order one.

3.4.2 General Autoregressive-Moving Average Process

In general, an *ARMA process* may consist of p AR parameters and q MA parameters. Such a process is denoted by ARMA(p,q) and is written as

$$(z_t - \mu) - \phi_1(z_{t-1} - \mu) - \phi_2(z_{t-2} - \mu) - \cdots - \phi_p(z_{t-p} - \mu)$$
$$= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad [3.4.3]$$

By implementing the B operator, [3.4.3] can be presented more conveniently as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(z_t - \mu) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

or

$$\phi(B)(z_t - \mu) = \theta(B)a_t \qquad [3.4.4]$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the AR operator of order p and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the MA operator of order q.

As mentioned by Box and Jenkins (1976, p. 74), an ARMA(p,q) process may be interpreted in two ways. It can be considered as a pth order AR process given by

$$\phi(B)(z_t - \mu) = e_t \qquad [3.4.5]$$

where e_t follows the qth order MA process

$$e_t = \theta(B)a_t \qquad [3.4.6]$$

Alternatively, an ARMA(p,q) process can be thought of as a qth order MA process

$$(z_t - \mu) = \theta(B)b_t \qquad [3.4.7]$$

where b_t follows the pth order AR process

$$\phi(B)b_t = a_t \qquad [3.4.8]$$

By substituting either e_t from [3.4.6] into [3.4.5] or else b_t from [3.4.8] into [3.4.7], it follows that

$$\phi(B)(z_t - \mu) = \theta(B)a_t$$

The ARMA(p,q) process contains both the pure AR and MA processes as subsets. Consequently, an AR(p) process is equivalent to an ARMA(p,0) process while a MA(q) process is the same as an ARMA(0,q) process. The ARMA(p,q) family of processes are also sometimes referred to as stationary nonseasonal Box-Jenkins processes because of the comprehensive presentation of these models in the book by Box and Jenkins (1976).

Stationarity and Invertibility

The conditions regarding stationarity and invertibility for AR and MA processes, also hold for ARMA processes. In order for an ARMA(p,q) process to be stationary the roots of the characteristic equation $\phi(B) = 0$ must fall outside the unit circle. Similarly, the roots of $\theta(B) = 0$ must fall outside the unit circle if the process is invertible and can be expressed as a pure AR process.

Autocorrelation Function

The theoretical ACF for an ARMA(p,q) process is derived in a fashion which is similar to that used for an AR process in Section 3.2.2. Multiply both sides of [3.4.3] by $(z_{t-k} - \mu)$ and take expectations to obtain

$$\begin{aligned}&\gamma_k - \phi_1\gamma_{k-1} - \phi_2\gamma_{k-2} - \cdots - \phi_p\gamma_{k-p}\\ &\quad = \gamma_{za}(k) - \theta_1\gamma_{za}(k-1) - \theta_2\gamma_{za}(k-2) - \cdots - \theta_q\gamma_{za}(k-q)\end{aligned} \qquad [3.4.9]$$

where $\gamma_k = E[(z_{t-k} - \mu)(z_t - \mu)]$ is the theoretical autocovariance function and $\gamma_{za}(k) = E[(z_{t-k} - \mu)a_t]$ is the cross covariance function between z_{t-k} and a_t. Since z_{t-k} is dependent only upon the shocks which have occurred up to time $t-k$, it follows that

$$\begin{aligned}&\gamma_{za}(k) = 0 \quad , \; k > 0\\ &\gamma_{za}(k) \neq 0 \quad , \; k \leq 0\end{aligned} \qquad [3.4.10]$$

Because of the $\gamma_{za}(k)$ terms in [3.4.9], it is necessary to derive other relationships before it is possible to solve for the autocovariances. This can be effected by multiplying [3.4.3] by a_{t-k} and taking expectations to get

$$\begin{aligned}&\gamma_{za}(-k) - \phi_1\gamma_{za}(-k+1) - \phi_2\gamma_{za}(-k+2) - \cdots - \phi_p\gamma_{za}(-k+p)\\ &\quad = -[\theta_k]\sigma_a^2\end{aligned} \qquad [3.4.11]$$

where

$$[\theta_k] = \begin{cases}\theta_k & , \; k = 1,2,\ldots,q\\ -1 & , \; k = 0\\ 0 & , \; \mathit{otherwise}\end{cases}$$

and $E[a_{t-k}a_t]$ is defined in [3.2.2].

Equations [3.4.9] and [3.4.11] can be employed to solve for the theoretical *autocovariance function for an ARMA(p,q) process*. For $k > q$, [3.4.9] reduces to

$$\gamma_k - \phi_1\gamma_{k-1} - \phi_2\gamma_{k-2} - \cdots - \phi_p\gamma_{k-p} = 0$$

or

$$\phi(B)\gamma_k = 0 \qquad [3.4.12]$$

If $k > r = \max(p,q)$, [3.4.12] may be used to calculate the γ_k directly from the previous values. For $k = 0,1,2,\ldots,r$, use [3.4.11] to solve for the cross covariances, $\gamma_{za}(k)$, and then substitute the $\gamma_{za}(k)$ into [3.4.9]. By employing the algorithm of McLeod (1975) outlined in Appendix A3.2, the resulting equations can be solved to determine the theoretical autocovariance function for any ARMA(p,q) process where the values of the parameters are known. The theoretical ACF can then be determined by simply dividing by the variance.

By dividing [3.4.12] by γ_0, the difference equation for the theoretical *ACF for an ARMA(p,q) process* is

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\rho_k = \phi(B)\rho_k = 0 \quad , \ k > q \qquad [3.4.13]$$

Except for the fact that [3.4.13] is only valid beyond lag q, the equation is identical to [3.2.9] which is the theoretical ACF for an AR(p) process. Hence, the attenuating behaviour of the ACF beyond lag q for an ARMA(p,q) process is due to the AR component in the model and the starting values for the difference equation. If $q - p < 0$ the entire theoretical ACF, ρ_j, for $j = 0,1,2,\ldots$, will be composed of a mixture of damped exponential and/or damped sine waves which possess characteristics controlled by $\phi(B)$ and the starting values. When $q - p \geq 0$ the $q - p + 1$ initial values $\rho_0, \rho_1, \rho_2, \ldots, \rho_{q-p}$ will not follow this pattern. Furthermore, because of the structure of [3.4.9], the autocorrelations $\rho_1, \rho_2, \ldots, \rho_q$, are a function of both the MA and AR parameters.

Partial Autocorrelation Function

As a result of the MA operator, the ARMA(p,q) process in [3.4.4] can be written as an infinite AR process given by

$$a_t = \theta(B)^{-1}\phi(B)(z_t - \mu) \qquad [3.4.14]$$

where $\theta(B)^{-1}$ is an infinite series in B. Since the definition of the PACF is based upon an AR process, the theoretical PACF is infinite in extent and attenuates with increasing lag. At higher lags, the behaviour of the PACF depends upon the MA parameters and is dominated by a mixture of damped exponentials and/or damped sine waves.

Douglas Fir Tree Ring Data: Because both the ACF and PACF die off for an ARMA(p,q) process, it is sometimes difficult to determine which type of ARMA model to fit to a given data set. Often, it is necessary to study two or three tentative ARMA models. For instance, consider the time series of 700 tree ring indices for Douglas fir at the Navajo National Monument in Arizona. This data is available from 1263 to 1962 and is listed in a report by Stokes et al. (1973). The plots of the sample ACF and PACF are displayed in Figures 3.4.1 and 3.4.2, respectively, along with the 95% confidence limits. Because both plots seem to attenuate, it may be appropriate to fit some type of ARMA(p,q) model to the data. The large values of both the ACF and PACF at lag one indicate that perhaps an ARMA(1,1) model may adequately model the data, although other ARMA models should perhaps also be examined. Using the estimator described in Appendix A6.1, the estimated ARMA(1,1) model is

$$(1-0.682B)(z_t-99.400)=(1-0.424B)a_t \qquad [3.4.15]$$

This calibrated model satisfies the diagnostic checks described in Chapter 7.

By using the parameter values for the tree ring model given in [3.4.15] as input to equations [3.4.9] and [3.4.11], the theoretical ACF can be calculated. The theoretical ACF for the tree ring model shown in Figure 3.4.3 is statistically similar to the sample ACF in Figure 3.4.1. To calculate the theoretical PACF using Pagano's algorithm in Appendix A3.1, the values of the theoretical ACF are substituted for the ρ_k's in [3.2.17]. The theoretical PACF in Figure 3.4.4 has the same form as the sample PACF in Figure 3.4.2. Because the fitted ARMA model appears to statistically preserve both the historical ACF and PACF, this fact enhances the desirability of ARMA models for use in the natural sciences. Additionally, Section 3.6 explains why ARMA models are suitably designed for capturing the physical chracteristics of annual streamflows.

ARMA(1,1) Process

The ARMA(1,1) process is given in [3.4.2]. As is the case for the AR(1) process (see Section 3.2.2), in order for the ARMA(1,1) process to be stationary, $|\phi_1| < 1$. Similarly, because the MA(1) process is invertible if $|\theta_1| < 1$ (see Section 3.3.2), the ARMA(1,1) process is invertible when the same conditions are placed upon θ_1.

To derive the autocovariance function for an ARMA(1,1) model, first use [3.4.9] to obtain

$$\gamma_0=\phi_1\gamma_1+\sigma_a^2-\theta_1\gamma_{za}(-1)$$

$$\gamma_1=\phi_1\gamma_0-\theta_1\sigma_a^2$$

$$\gamma_k=\phi_1\gamma_{k-1}\ ,\quad k\geq 2$$

Next, after setting $k=1$, employ [3.4.11] to get

$$\gamma_{za}(-1)=(\phi_1-\theta_1)\sigma_a^2$$

where $\gamma_{za}(0)=\sigma_a^2$ in both [3.4.9] and [3.4.11]. Upon substituting $\gamma_{za}(-1)$ into the previous equation for γ_0, the autocovariances for an ARMA(1,1) process are found to be

$$\gamma_0=\frac{1+\theta_1^2-2\phi_1\theta_1}{1-\phi_1^2}\sigma_a^2$$

$$\gamma_1=\frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1-\phi_1^2}\sigma_a^2$$

$$\gamma_k=\phi_1\gamma_{k-1}\ ,\quad k\geq 2 \qquad [3.4.16]$$

By dividing by γ_0, the theoretical ACF of an ARMA(1,1) process is

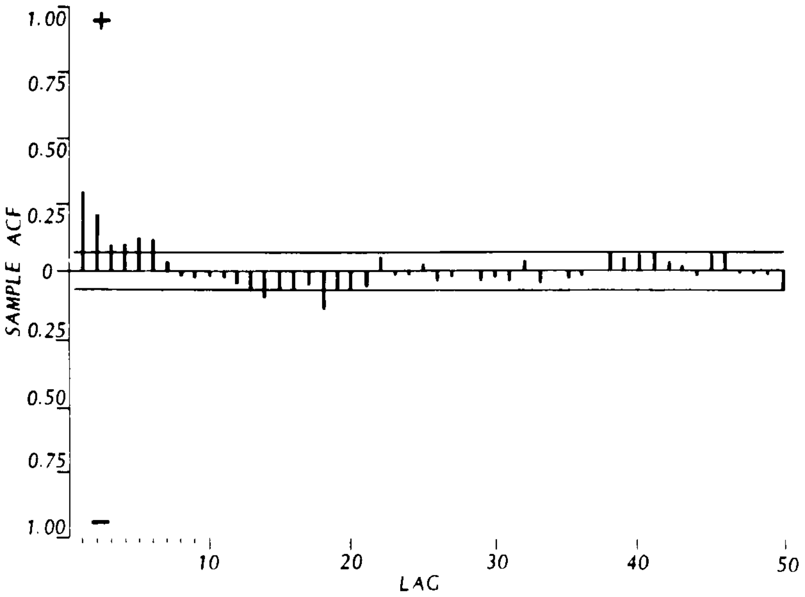


Figure 3.4.1. Sample ACF and 95% confidence limits for the Douglas Fir tree ring series at Navajo National Monument in Arizona.

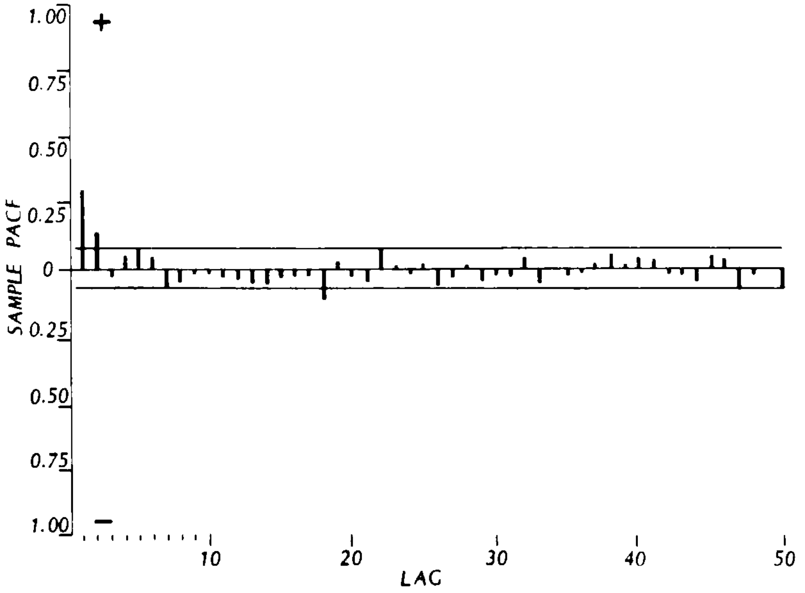


Figure 3.4.2. Sample PACF and 95% confidence limits for the Douglas Fir tree ring series at Navajo National Monument in Arizona.

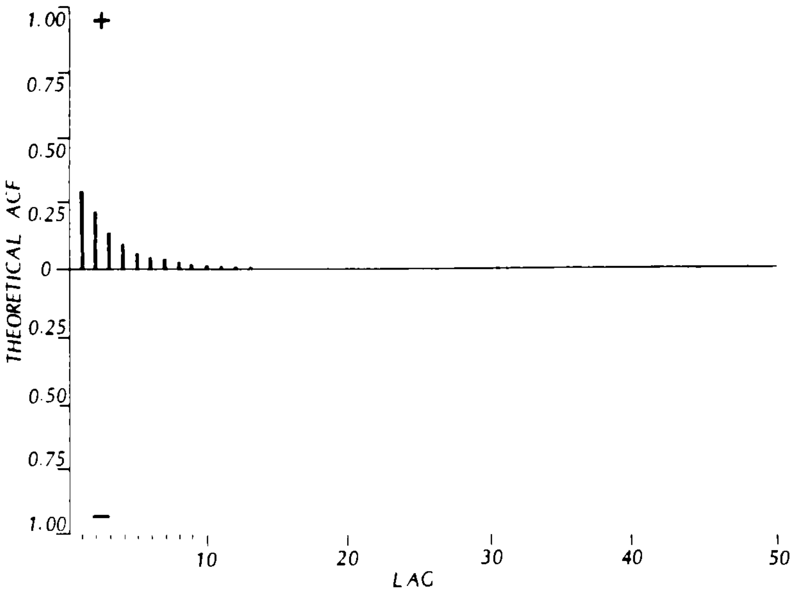


Figure 3.4.3. Theoretical ACF for the ARMA(1,1) model fitted to the Douglas Fir tree ring series at Navajo National Monument in Arizona.

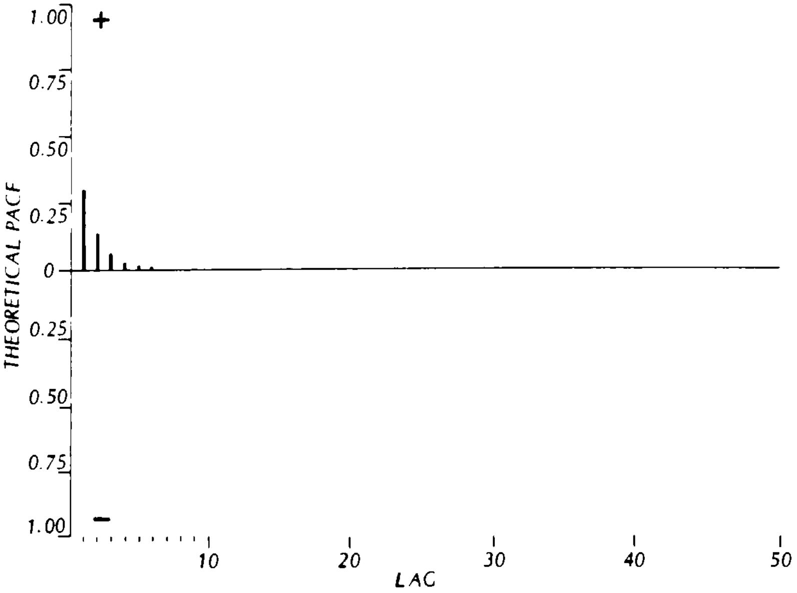


Figure 3.4.4. Theoretical PACF for the ARMA(1,1) model fitted to the Douglas Fir tree ring series at Navajo National Monument in Arizona.

$$\rho_1 = \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$$

$$\rho_k = \phi_1\rho_{k-1}\ ,\quad k \geq 2 \qquad [3.4.17]$$

From [3.4.17], it can be seen that ρ_1 is a function of both the MA and AR parameters. The ACF at lag 2 depends only upon ϕ_1 and also the starting value ρ_1. Furthermore, the theoretical ACF decays exponentially from the starting value ρ_1. This exponential decay is even when ϕ_1 is positive and is oscillatory whenever ϕ_1 is negative. In addition, the sign of $(\phi_1 - \theta_1)$ dictates the sign of ρ_1 and also from which side of zero the exponential decay commences.

By substituting the theoretical ACF in [3.4.17] into the Yule-Walker equations which are given in [3.2.17], the theoretical PACF can be determined for the ARMA(1,1) process. At lag 1, $\phi_{11} = \rho_1$, while for lags greater than one the PACF of an ARMA(1,1) process behaves like the PACF of a MA(1) process (see Section 3.3.2) and hence follows the form of a damped exponential. When θ_1 is positive, the PACF consists of an evenly damped exponential which decays from ρ_1, where the sign of ρ_1 is determined by the sign of $(\phi_1 - \theta_1)$. If θ_1 is negative, the PACF is dominated by an oscillating exponential which attenuates from $\phi_{11} = \rho_1$, where the sign of ρ_1 is determined by $(\phi_1 - \theta_1)$.

3.4.3 Three Formulations of the Autoregressive-Moving Average Process

An ARMA(p,q) process can be expressed in three explicit forms. One formulation is to use the difference equation given in [3.4.4]. A second method is to express the process as a pure MA process. This is also referred to as the random shock form of the process. Finally, the third option is to formulate the process as a pure AR process which is also called the inverted form of the process.

Random Shock Form

Because $\phi(B)$ and $\theta(B)$ can be treated as algebraic operators, the ARMA(p,q) process can be written in *random shock form* as

$$\begin{aligned}(z_t - \mu) &= \phi(B)^{-1}\theta(B)a_t \\ &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= a_t + \psi_1 B a_t + \psi_2 B^2 a_t + \cdots \\ &= (1 + \psi_1 B + \psi_2 B^2 + \ldots)a_t \\ &= \psi(B)a_t \qquad [3.4.18]\end{aligned}$$

where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$, is the *random shock or infinite MA operator* and ψ_i is the ith parameter, coefficient or weight of $\psi(B)$. It is often convenient to express an ARMA process in the form given in [3.4.18] for both theoretical and application purposes. For instance, the ψ weights are required in Section 8.2.3 to calculate the variance of the forecasts. As explained in Section 9.3, one way to simulate data is to first express an ARMA model in random shock form and then use this format of the model for simulation purposes. By writing each member of a set of ARMA models in random shock form, the models can be conveniently compared by looking at the magnitude and sign of the ψ parameters. Furthermore, nonstationary processes (see Chapter 4) and seasonal processes (see Part VI) can also be written in the general random shock form of the process.

As is noted in Section 3.2.2, if an AR(p) process is stationary the roots of $\phi(B)=0$ must lie outside the unit circle and this insures that the process can also be written as an infinite MA process which will converge for $|B|\leq 1$. Consequently, a necessary condition for stationarity for an ARMA(p,q) process, is that the weights $\psi_1,\psi_2,\ldots$, in $\psi(B)=\phi(B)^{-1}\theta(B)$, form a convergent series for $|B|\leq 1$ [see Box and Jenkins (1976, Appendix A3.1, pp. 80-82) for a mathematical proof]. The stationarity requirement is proven by examining the theoretical autocovariance which is given by

$$\gamma_k=\sigma_a^2\sum_{j=0}^{\infty}\psi_j\psi_{j+k} \quad , \quad k=0,1,2,\ldots \tag{3.4.19}$$

By substituting $k=0$ into [3.4.19], the variance is found to be

$$\gamma_0=\sigma_z^2=\sigma_a^2\sum_{j=0}^{\infty}\psi_j^2 \tag{3.4.20}$$

In order to have a finite variance and hence stationarity, the ψ weights must decrease in a manner which allows the right side of [3.4.20] to converge.

To develop a relationship for determining the ψ parameters or weights, first multiply [3.4.18] by $\phi(B)$ to obtain

$$\phi(B)(z_t-\mu)=\phi(B)\psi(B)a_t$$

From [3.4.4], $\theta(B)a_t$ can be substituted for $\phi(B)(z_t-\mu)$ in the previous equation to get

$$\phi(B)\psi(B)=\theta(B)$$

The ψ weights can be conveniently determined by expressing the above equation as

$$\phi(B)\psi_k=-\theta_k \tag{3.4.21}$$

where B operates on k, $\psi_0=1$, $\psi_k=0$ for $k<0$, and $\theta_k=0$ if $k>q$. When $k>q$ then $\theta_k=0$ and ψ_k in [3.4.21] satisfies the same difference equation as the theoretical ACF of an AR process and also an ARMA process in [3.2.9] and [3.4.13], respectively. Consequently, when ψ_k is plotted against lag k it will follow the same type of patterns as the theoretical ACF of the process. For increasing lags, the graph may follow a smooth exponential decay, an exponential decline that alternates in sign, or a damped sinusoidal decay.

Given the AR and MA parameters, one can employ [3.4.21] to calculate the random shock parameters. To decide upon how many ψ coefficients to estimate, one can calculate enough ψ coefficients to keep the relative error in the variance of the random shock process less than a specified error level. When σ_a^2 is assumed to be one and the ψ's are only considered up to lag q', from [3.4.20] the variance of the random shock process is approximately given by

$$\sum_{j=0}^{q'} \psi_j^2$$

When it is assumed that $\sigma_a^2 = 1$, one can calculate the variance, γ_0, of the given ARMA(p,q) process by solving [3.4.9] and [3.4.11]. Consequently, the relative absolute error due to the random shock approximation is

$$\left| \frac{\gamma_0 - \sum_{j=0}^{q'} \psi_j^2}{\gamma_0} \right|$$

One can choose q' to be just large enough to cause the above expression to have a value less than the specified error level. To demonstrate how the ψ coefficients are calculated using [3.4.21], two examples are now given.

Example Using the Temperature Model: In Section 3.3.2 it is noted that it may be appropriate to fit either a MA(2) or an AR(2) model to the annual temperature data from the English Midlands. Because the sample ACF in Figure 2.5.1 seems to truncate after lag 2, a MA(2) model may be needed. However, since the sample PACF in Figure 3.3.1 cuts off after lag 2, an AR(2) model may be suitable for modelling the series. In reality, either of these models may be employed, since they are quite similar. This can be demonstrated by expanding the AR(2) model as an infinite MA model and then comparing the results to the MA(2) model in [3.3.12].

The estimated AR(2) model for the temperature data is

$$(1 - 0.119B - 0.200B^2)(z_t - 9.216) = a_t \qquad [3.4.22]$$

where 9.216 is the MLE of the mean level. For the model in the above equation, [3.4.21] becomes

$$(1 - 0.119B - 0.200B^2)\psi_k = 0$$

When $k = 1$

$$(1 - 0.119B - 0.200B^2)\psi_1 = 0 \text{ or } \psi_1 - 0.119\psi_0 - 0.200\psi_{-1} = 0$$

Since $\psi_0 = 1$ and $\psi_{-1} = 0$ the expression reduces to

$$\psi_1 = 0.119$$

For $k = 2$

$$(1 - 0.119B - 0.200B^2)\psi_2 = 0 \text{ or } \psi_2 - 0.119\psi_1 - 0.200\psi_0 = 0$$

Therefore, $\psi_2 = 0.119(0.119) + 0.200 = 0.214$.

When $k = 3$

$$(1 - 0.119B - 0.200B^2)\psi_3 = 0 \text{ or } \psi_3 - 0.119\psi_2 - 0.200\psi_1 = 0$$

Therefore, $\psi_3 = 0.119(0.214) + 0.200(0.119) = 0.049$.

In general, the expression for ψ_k is

$$\psi_k = 0.119\psi_{k-1} + 0.200\psi_{k-2},\ k > 0$$

Because of the form of this equation, ψ_k decays towards zero rather quickly for increasing lag after lag 2.

Using the results for the ψ coefficients, the random shock form of the AR(2) model in [3.4.22] is

$$\begin{aligned} z_t - 9.216 &= (1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots)a_t \\ &= (1 + 0.119B + 0.214B^2 + 0.049B^3 + \cdots)a_t \end{aligned} \quad [3.4.23]$$

The SE of estimation for both MA parameters in [3.3.12] is 0.062 and it can be seen that each MA parameter in [3.4.23] is within one SE of the corresponding MA parameter in [3.3.12]. Consequently, for practical purposes the AR(2) model in [3.4.22] is actually the same as the MA(2) model in [3.3.12].

Example Using the Tree Ring Model: The sample ACF and PACF are shown in Figures 3.4.1 and 3.4.2, respectively, for the Douglas Fir tree ring series at Navajo National Monument in Arizona. Because both of these plots attenuate, it may be appropriate to fit an ARMA(1,1) model to this series. The fitted model for this data is given in [3.4.15].

For the ARMA(1,1) model, [3.4.21] becomes

$$(1 - 0.682B)\psi_k = -\theta_k$$

where $\theta_k = 0$ for $k > 1$. When $k = 1$

$$(1 - 0.682B)\psi_1 = -0.424 \quad \text{or} \quad \psi_1 - 0.682\psi_0 = -0.424$$

Therefore, $\psi_1 = 0.682 - 0.424 = 0.258$.

For $k = 2$

$$(1 - 0.682B)\psi_2 = 0 \quad \text{or} \quad \psi_2 - 0.682\psi_1 = 0$$

Hence, $\psi_2 = 0.682\psi_1 = 0.682(0.258) = 0.176$.

When $k = 3$

$$(1 - 0.682B)\psi_3 = 0 \quad \text{or} \quad \psi_3 - 0.682\psi_2 = 0$$

Therefore, $\psi_3 = 0.682\psi_2 = 0.682(0.176) = 0.120$.

The general expression for ψ_k is

$$\psi_k = 0.682\psi_{k-1} = (0.682)^{k-1}\psi_1 \quad , \ k > 0$$

Due to the form of this equation, ψ_k will decrease in absolute value for increasing lag. When the ARMA(1,1) model is expressed using the ψ coefficients, the random shock form of the models is

$$z_t - 99.400 = (1 + 0.258B + 0.176B^2 + 0.120B^3 + ...)a_t \qquad [3.4.24]$$

Inverted Form

To express the ARMA(p,q) process in *inverted form* as a pure AR process, [3.4.4] can be written as

$$\begin{aligned} a_t &= \theta(B)^{-1}\phi(B)(z_t - \mu) \\ &= (z_t - \mu) - \pi_1(z_{t-1} - \mu) - \pi_2(z_{t-2} - \mu) - \cdots \\ &= (z_t - \mu) - \pi_1 B(z_t - \mu) - \pi_2 B^2(z_t - \mu) - \cdots \\ &= (1 - \pi_1 B - \pi_2 B^2 - ...)(z_t - \mu) \\ &= \pi(B)(z_t - \mu) \qquad [3.4.25] \end{aligned}$$

where $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$, is the *inverted or infinite AR operator* and π_i is the ith parameter, coefficient or weight of $\pi(B)$. Besides ARMA models, it is often convenient to write nonstationary processes and also various types of seasonal processes in the above format. Furthermore, by comparing [3.4.18] and [3.4.25] it is evident that

$$\psi(B)^{-1} = \pi(B) \qquad [3.4.26]$$

In Section 3.3.2, it is pointed out that in order for a MA(q) process to be expressed as a meaningful infinite AR process the roots of $\theta(B) = 0$ must lie outside the unit circle. Invertibility is also achieved for an ARMA(p,q) process when the roots of $\theta(B) = 0$ lie outside the unit circle. This is because the weights $\pi_1, \pi_2, ...,$ in the inverted operator $\pi(B) = \theta(B)^{-1}\phi(B)$ constitute a convergent series for $|B| \leq 1$. The invertibility condition is independent of the stationarity condition and can also be used with nonstationary processes.

To determine a relationship for computing the π parameters, multiply [3.4.25] by $\theta(B)$ to get

$$\theta(B)a_t = \theta(B)\pi(B)(z_t - \mu)$$

Using [3.4.4], $\phi(B)(z_t - \mu)$ can be substituted for $\theta(B)a_t$ in the above equation to obtain

$$\phi(B) = \theta(B)\pi(B)$$

The π weights can be readily ascertained by expressing the above equation as

$$\theta(B)\pi_k = \phi_k \qquad [3.4.27]$$

where B operates on k, $\pi_0 = -1$ when using [3.4.27] to calculate π_k for $k > 0$, $\pi_k = 0$ for $k < 0$, and $\phi_k = 0$ if $k > p$. When $k > p$, π_k satisfies the same difference equation as the inverse autocorrelation function (IACF) that is discussed in Section 5.3.6. Consequently, when π_k is plotted against lag k it will possess the same behaviour as the IACF. For increasing lags, the graph may consist of a smooth exponential decay, an exponential decline where the values alternate in sign or a damped sinusoidal decay. Some examples are now presented to demonstrate how to employ [3.4.27] for calculating the π parameters by hand.

Example Using the Temperature Model: The MA(2) model in [3.3.12] for the average annual temperature data from the English Midlands can be equivalently expressed as an infinite AR model. To determine the π weights for the MA(2) model, [3.4.27] becomes

$$(1+0.111B+0.197B^2)\pi_k=0$$

When $k=1$

$$(1+0.111B+0.197B^2)\pi_1=0 \quad \text{or} \quad \pi_1+0.111\pi_0+0.197\pi_{-1}=0$$

Since $\pi_0=-1$ and $\pi_{-1}=0$, the expression reduces to $\pi_1=0.111$.

For $k=2$

$$(1+0.111B+0.197B^2)\pi_2=0 \quad \text{or} \quad \pi_2+0.111\pi_1+0.197\pi_0=0$$

Therefore, $\pi_2=-0.111(0.111)+0.197=0.185$.

When $k=3$

$$(1+0.111B+0.197B^2)\pi_3=0 \quad \text{or} \quad \pi_3+0.111\pi_2+0.197\pi_1=0$$

Hence, $\pi_3=-0.111(0.185)-0.197(0.111)=-0.042$.

In general, the expression for π_k is

$$\pi_k=-0.111\pi_{k-1}-0.197\pi_{k-2}, \quad k>0$$

Because of the structure of the above equation, π_k attenuates quickly in absolute value after lag 2.

By employing the results for the π weights, the inverted form of the MA(2) model in [3.3.12] is

$$(1-0.111B-0.185B^2+0.042B^3+\ldots)(z_t-9.216)=a_t \qquad [3.4.28]$$

It can be seen that inverted form of the MA(2) model in [3.4.28] is almost the same as the AR(2) model in [3.4.22] for the temperature data. The SE of estimation for both AR parameters in [3.4.22] is 0.062 and each AR parameter in [3.4.28] is within one SE of the corresponding AR parameter in [3.4.22]. This confirms that the MA(2) model in [3.3.12] is statistically the same as the AR(2) model in [3.4.22].

Example Using the Tree Ring Model: The ARMA(1,1) model which is fitted to the Douglas Fir tree ring series at Navajo National Monument in Arizona is given in [3.4.15]. This model can be equivalently expressed as an inverted model by using [3.4.27] to obtain the π weights. For the case of the ARMA(1,1) model, [3.4.27] becomes

$$(1-0.424B)\pi_k=\phi_k$$

where $\phi_k=0$ for $k>1$ and $\pi_0=-1$.

When $k=1$

$$(1 - 0.424B)\pi_1 = 0.682 \text{ or } \pi_1 - 0.424\pi_0 = 0.682$$

Therefore, $\pi_1 = -0.424 + 0.682 = 0.258$.

For $k = 2$

$$(1 - 0.424B)\pi_2 = 0 \text{ or } \pi_2 - 0.424\pi_1 = 0$$

Hence, $\pi_2 = 0.424\pi_1 = 0.424(0.258) = 0.109$.

When $k = 3$

$$(1 - 0.424B)\pi_3 = 0 \text{ or } \pi_3 - 0.424\pi_2 = 0$$

Therefore, $\pi_3 = 0.424\pi_2 = 0.424(0.109) = 0.046$.

The general expression for π_k is

$$\pi_k = 0.424\pi_{k-1} = (0.424)^{k-1}\pi_1, \quad k > 0$$

It can be seen from this equation that π_k will decrease in absolute value for increasing lag. When the ARMA(1,1) model is written using the π parameters, the inverted form of the model is

$$(1 - 0.258B - 0.109B^2 - 0.046B^3 - \cdots)(z_t - 99.40) = a_t \qquad [3.4.29]$$

Linear Filter Interpretation

The random shock form of the process in [3.4.18] can be considered in terms of a linear filter. As shown in Figure 3.4.5, the white noise input passes through the *linear filter* $\psi(B)$ which transforms the white noise into the output $(z_t - \mu)$. Because of this, the random shock operation $\psi(B)$ is referred to as the transfer function or the filter. When the sequence formed by the ψ weights is either finite or infinite and convergent, the filter is stable because the process z_t is stationary. For stationary processes μ is the mean level about which the process varies. However, when the filter is unstable and the process is not stationary, by definition the process does not fluctuate about any mean level and μ can be considered as a reference point.

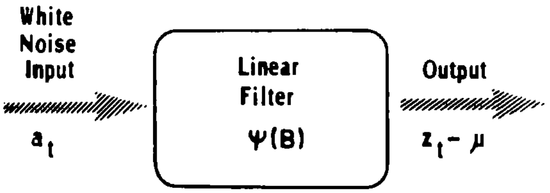


Figure 3.4.5. Linear filter interpretation of the random shock model.

Linear Difference Equations

Equation [3.4.4] for an ARMA(p,q) process constitutes what is called a *linear difference equation* because the process is linear with respect to the AR and MA parameters. Similarly, the random shock and inverted forms of the ARMA model in [3.4.18] and [3.4.25], respectively, are also linear difference equations. Another example of a linear difference equation is [3.2.9] for the theoretical ACF of an AR(p) process. Difference equations arise in time series modelling because it is necessary to model time series which have values at discrete and evenly spaced time intervals. On the other hand, differential equations are employed for modelling systems which evolve over continuous time.

The solution of a linear difference equation is analogous to that for a *linear differential equation*. The final solution for a linear differential equation consists of an equation which does not possess any differentials. Similarly, the solution to a linear difference equation is an equation which does not contain entries which can be written in terms of the B operator. As is the case for a linear differential equation, the general solution for a linear difference equation consists of the summation of a complementary function plus a particular integral. For a brief description of how to solve linear difference equations, the reader may wish to refer to Box and Jenkins (1976, Appendix A4.1, pp. 114-119). Pandit and Wu (1983) make extensive comparisons between linear stochastic differential and difference equations. In fact, these authors explain how to obtain both difference and differential equations from a time series to represent the underlying dynamic system and how to employ these equations for prediction, control and other applications.

3.4.4 Constrained Models

As mentioned earlier, a primary objective in stochastic modelling is to adequately model the data using a model which possesses as few parameters as possible. The principle of model parsimony can be achieved in practice by using a discrimination procedure such as the Akaike information criterion (Akaike, 1974) (see Sections 1.3.3 and 6.3) in conjunction with diagnostic checks (see Chapter 7). This can result in selecting an ARMA(p,q) model where some of the AR and MA parameters which are less than order p and q respectively, are omitted from the model. For instance, as shown in Section 3.2.2, the most appropriate model to fit to the average annual flows of the St. Lawrence River at Ogdensburg, New York is an AR(3) model without the ϕ_2 parameter. The difference equation for this model is given in [3.2.19]. Models which have some of the parameters constrained to zero are referred to as *constrained models*. The option for omitting model parameters can be done with both nonseasonal and seasonal models.

An interesting constrained model is the one which is fitted to the yearly Wolfer sunspot number series in Section 6.4.3. This sunspot series is available from 1700 to 1960 in the work of Waldmeier (1961). If it is deemed appropriate to fit an ARMA model to the sunspot series, it turns out that the best ARMA model is an AR(9) model with ϕ_3 to ϕ_8 left out of the model. In addition, as is shown in Section 6.4.3, it is first necessary to take a square root transformation of the data before fitting the constrained AR(9) model.

A constrained AR model is also referred to as a subset AR model. Research on this topic is provided by authors such as Haggan and Oyetunji (1984) as well as Yu and Lin (1991). Moreover, subset autoregression is also discussed in Section 6.3.6.

3.4.5 Box-Cox Transformation

As noted in Section 3.2.1, the a_t series is always assumed to be independently distributed and possess a constant variance about a zero mean level. In addition, it is often appropriate to invoke the normality assumption for the residuals in order to obtain MLE's for the model parameters (Chapter 6) and subsequently to carry out diagnostic checks (Chapter 7). When fitting ARMA models to a given data set, the model residuals can be estimated along with the model parameters at the estimation stage and model adequacy can be ascertained by checking that the residual assumptions are satisfied. The independence assumption is the most important of all and its violation can cause drastic consequences (Box and Tiao, 1973, p. 522). In fact, when the independence assumption is violated it is necessary to design another model to fit to the data (see Chapter 7). However, if the constant variance and/or normality assumptions are not true, they are often reasonably well fulfilled when the observations are transformed by a Box-Cox transformation (Hipel et al., 1977; McLeod et al., 1977).

A *Box-Cox transformation* (Box and Cox, 1964) is defined by

$$z_t^{(\lambda)} = \begin{cases} \lambda^{-1}[(z_t + c)^{\lambda} - 1] & , \ \lambda \neq 0 \\ ln(z_t + c) & , \ \lambda = 0 \end{cases} \qquad [3.4.30]$$

where c is a constant. The power transformation in [3.4.30] is valid for $z_t + c > 0$. Consequently, if all of the values in the time series are greater than zero usually the constant is set equal to zero. When negative and/or zero values of z_t are present it is usually most convenient to select the constant to be slightly larger than the absolute value of the largest non-positive entry in the time series.

Because the parameter values of an ARMA model fitted to a given time series are unchanged by a linear transformation, the transformation in [3.4.30] is equivalent to

$$z_t^{(\lambda)} = \begin{cases} z_t^{\lambda} & , \ \lambda \neq 0 \\ lnz_t & , \ \lambda = 0 \end{cases} \qquad [3.4.31]$$

where the entries of the z_t series are all greater than zero. The form of the Box-Cox transformation in [3.4.30] is preferable theoretically to that in [3.4.31] because the transformation in [3.4.30] is continuous at $\lambda = 0$. By invoking L'Hopital's rule, it can be shown that the transformation for $\lambda \neq 0$ in [3.4.30] reduces to $ln(z_t + c)$ in the limit as λ approaches zero. When $\lambda = 1$, this means that there is no power transformation.

After the entries in a time series have been changed by a transformation such as that given in [3.4.30], or others discussed by Jain and Singh (1986), an appropriate ARMA model can be fitted to the transformed data. The equation for an ARMA(p,q) model for the $z_t^{(\lambda)}$ series is

$$\phi(B)(z_t^{(\lambda)} - \mu) = \theta(B)a_t \qquad [3.4.32]$$

where μ is the mean level of the $z_t^{(\lambda)}$ sequence. Box-Cox transformations are useful when dealing with both nonseasonal and seasonal time series. For notational convenience in later chapters often $z^{(\lambda)}$ is simply written as z_t where it is assumed that the series to which the model is fitted is transformed using [3.4.30] whenever necessary. Finally, data transformations that can be considered when dealing with extreme values are referred to in Section 5.3.3.

3.5 THEORETICAL SPECTRUM

As noted in Section 2.6, most of the time series modelling and analysis methods presented in this book are defined and used within the time domain. For example, the theoretical ACF and PACF for an ARMA model constitute time domain properties which are needed for model identification in Chapter 5. Based upon a knowledge of the general properties of the theoretical ACF and PACF, one can examine the characteristics of the sample ACF and PACF for deciding upon which parameters to include in an ARMA model to fit to a given data set.

The objective of this section is to define the theoretical spectrum for ARMA models and present some graphs of the spectrum for specific kinds of ARMA models. As explained below, the spectral density is simply the Fourier transform of the theoretical autocovariance function. Consequently, the spectral density is simply the representation of the autocovariance function within the frequency domain.

3.5.1 Definitions

Any stationary time series, z_t, can be viewed as being composed of a limiting sum of sinusoids of the form

$$A_i\cos(2\pi f_i t + \alpha_i)$$

where f_i is the frequency, A_i is the amplitude and α_i is the phase. The frequency varies from -1/2 to 1/2 in cycles per unit time. The amplitude and phase components at frequency f_i are uncorrelated random variables with a mean of zero in each different realization of the time series. The variance of the amplitude is determined by the spectrum which is defined in the next paragraph. Those frequencies for which the spectrum, $S(f)$, is large will contribute sinusoids with greater amplitudes and thus represent more important sources of variation in the time series.

The Cramer spectral representation expresses the aforesaid facts in a more precise fashion. Every covariance stationary time series with a mean of zero has the Cramer spectral representation [see for example, Kleiner et al. (1979, p. 319)]

$$z_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f) \qquad [3.5.1]$$

where $Z(f)$ for $|f| \le 1/2$ is a continuous stochastic process with orthogonal increments (so that $Z(f_2) - Z(f_1)$ and $Z(f_4) - Z(f_3)$ are uncorrelated whenever $f_1 < f_2 \le f_3 < f_4$). The process $Z(f)$ defines the cumulative spectral density function $F(f)$, by

$$F(f) = E\left[|Z(f)|^2\right]$$

$$dF(f) = E\left[|dZ(f)|^2\right]$$

$$F(-1/2) = 0$$

$$F(1/2) = var(z_t) = \gamma_0 \quad [3.5.2]$$

For most types of time series, the derivative of $F(f)$ exists and the *spectrum* may be defined as

$$S(f) = 2F'(f), \quad 0 \le f \le 1/2 \quad [3.5.3]$$

The factor of 2 on the right hand side of [3.5.3] allows for the fact that the spectrum is symmetric about zero and hence only the spectrum in the range $0 \le f \le 1/2$ needs to be considered. In addition to spectrum, other commonly used names for $S(f)$ are spectral density, power spectral density, spectral density function, power spectral density function and power spectrum.

It follows from [3.5.1] and the orthogonal increment property of $A(f)$ that

$$\gamma_k = 1/2 \int_{-1/2}^{1/2} e^{i2\pi fk} S(f)\, df$$

$$= 1/2 \int_{-1/2}^{1/2} (\cos 2\pi fk + i \sin 2\pi fk)\, S(f)\, df$$

$$= \int_{0}^{1/2} \cos 2\pi fk\; S(f)\, df \quad [3.5.4]$$

For $k = 0$,

$$\gamma_0 = \int_{0}^{1/2} S(f)\, df \quad [3.5.5]$$

Because of [3.5.5], the spectrum gives the distribution of the variance of the process over frequency and the area under the spectral curve is the variance.

By taking the inverse transformation of [3.5.4], it follows that the spectral density function is given by

$$S(f) = 2 \sum_{k=-\infty}^{\infty} \gamma_k \cos 2\pi fk \quad [3.5.6]$$

The above equation shows that the spectrum is simply the Fourier transform of the autocovariance function.

The spectrum can conveniently be written in terms of the autocovariance generating function. When an ARMA process is expressed as the random shock form of the process in [3.4.18], the *autocovariance generating function* is given as (Box and Jenkins, 1976, p. 81)

$$\gamma(B) = \sigma_a^2 \psi(B) \psi(B^{-1}) \quad [3.5.7]$$

where B^{-1} is the *forward shift operator* defined by

$$B^{-1} z_t = z_{t+1} \text{ and } B^{-k} z_t = z_{t+k}$$

Because the spectrum is the Fourier transform of the autocovariance function, it can be written in terms of the autocovariance generating function in [3.5.4] as

$$S(f) = 2\gamma(e^{i2\pi f}) = 2\sigma_a^2\psi(e^{i2\pi f})\psi(e^{-i2\pi f})$$
$$= 2\sigma_a^2|\psi(e^{-i2\pi f})|^2 \qquad [3.5.8]$$

When utilizing the AR and MA operators, [3.5.8] for an ARMA process is given as

$$S(f) = 2\sigma_a^2\left|\frac{\theta(e^{-i2\pi f})}{\phi(e^{-i2\pi f})}\right|^2 \qquad [3.5.9]$$

To calculate the theoretical spectrum for an ARMA process, the sum of angles method (Robinson, 1967, p. 64; Otnes and Enochson, 1972, p. 139) can be used to recursively calculate the sine and cosine terms (see [2.6.3] and [2.6.4]).

The *normalized spectral density function* is given by

$$s(f) = \frac{S(f)}{\gamma_0} \qquad [3.5.10]$$

Because $s(f)$ is not a function of σ_a^2, it is often used instead of $S(f)$. For the applications in this section, the normalized spectral density is employed.

Examples: Consider obtaining the autocovariance function and the normalized spectrum for a MA(1) process by employing [3.5.7] and [3.5.10], respectively. When using the autocovariance generating function to ascertain γ_k, the coefficient of either B^k or B^{-k} are examined in [3.5.7]. For a MA(1) process, $\psi(B) = 1 - \theta_1 B$ and the autocovariance generating function is

$$\gamma(B) = \sigma_a^2(1 - \theta_1 B)(1 - \theta_1 B^{-1})$$
$$= \sigma_a^2(-\theta_1 B^{-1} + (1 + \theta_1^2) - \theta_1 B)$$

From the coefficients of the backward shift operator, the autocovariances are found to be

$$\gamma_0 = (1 + \theta_1^2)\sigma_a^2$$
$$\gamma_1 = -\theta_1\sigma_a^2$$
$$\gamma_k = 0 \quad k \geq 2$$

By utilizing [3.5.9], the spectrum for a MA(1) process is

$$S(f) = 2\sigma_a^2|1 - \theta_1(e^{-i2\pi f})|^2$$
$$= 2\sigma_a^2[(1 - \theta_1\cos 2\pi f)^2 + (\theta_1\sin 2\pi f)^2]$$
$$= 2\sigma_a^2(1 - 2\theta_1\cos 2\pi f + \theta_1^2)$$

From [3.3.10], the variance of a MA(1) process is $(1 + \theta_1^2)\sigma_a^2$ and, consequently, the normalized spectrum is calculated using [3.5.10] as

$$s(f) = \frac{2(1 - 2\theta_1\cos 2\pi f + \theta_1^2)}{(1 + \theta_1^2)} \qquad [3.5.11]$$

3.5.2 Plots of the Log Normalized Spectrum

For a white noise process, the normalized spectrum in [3.5.10] reduces to

$$S(f) = 2 \qquad [3.5.12]$$

Consequently, all frequencies are equally important for explaining the process and a graph of $S(f)$ against frequency would simply be a straight line.

When considering an AR(1) process the variance of the process is given in [3.2.16] as $\sigma_a^2/(1-\phi_1^2)$. The normalized spectrum is calculated from [3.5.10] as

$$\begin{aligned} s(f) &= \frac{2(1-\phi_1^2)}{|1-\phi_1 e^{-i2\pi f}|^2} \\ &= \frac{2(1-\phi_1^2)}{(1-\phi_1\cos 2\pi f)^2 + (\phi_1 \sin 2\pi f)^2} \\ &= \frac{2(1-\phi_1^2)}{1-2\phi_1\cos 2\pi f + \phi_1^2} \end{aligned} \qquad [3.5.13]$$

For $\phi_1 > 0$, the normalized spectrum in [3.5.13] is easily seen to be a steadily decreasing function for increasing frequency. This means that most of the variance of the time series can be represented as low frequency sinusoids. When the natural logarithms of the normalized spectrum are plotted against frequency, this may improve the ability to distinguish important features of the graph. The log normalized spectrums for AR(1) processes with $\phi_1 = 0.3$ and $\phi_1 = 0.8$ are displayed in Figures 3.5.1 and 3.5.2, respectively. The spectrums are calculated at enough points to cause the curves to appear to be smooth. As can be seen, low frequencies are dominant in both of these figures and consequently the spectrums are said to be "red" (this is because red is on the low frequency end of visible light in the electromagnetic spectrum). Furthermore, because the process with $\phi_1 = 0.3$ is closer to white noise than the process with $\phi_1 = 0.8$, the log normalized spectrum in Figure 3.5.1 is "flatter" than the plot in Figure 3.5.2.

When ϕ_1 for an AR(1) process is negative, the spectrum is dominated by high frequencies. Figures 3.5.3 and 3.5.4 are plots of the log normalized spectrum for $\phi_1 = -0.3$ and $\phi_1 = -0.8$, respectively. As shown in these graphs, most of the variance is explained by high frequencies terms in the "blue" end of the frequency scale. In addition, the upswing in the log normalized spectrum in the high frequencies is more pronounced for the process with $\phi_1 = -0.80$ as compared to the case when $\phi_1 = -0.30$.

Other theoretical spectrums can be readily examined by employing [3.5.10]. Of particular interest are the spectrums of the models that have been fitted to various geophysical time series. This is because a plot of the spectral density of a fitted ARMA model can be useful in obtaining insight into important properties of the original time series.

Figure 3.5.5 shows a plot of the log normalized spectrum for the constrained AR(3) model without ϕ_2 that is fitted to the average annual flows of the St. Lawrence River at Ogdensburg, New York. The difference equation for this model is given in [3.2.19]. As can be observed in Figure 3.5.5, the low frequencies are most important for explaining the variance. From a

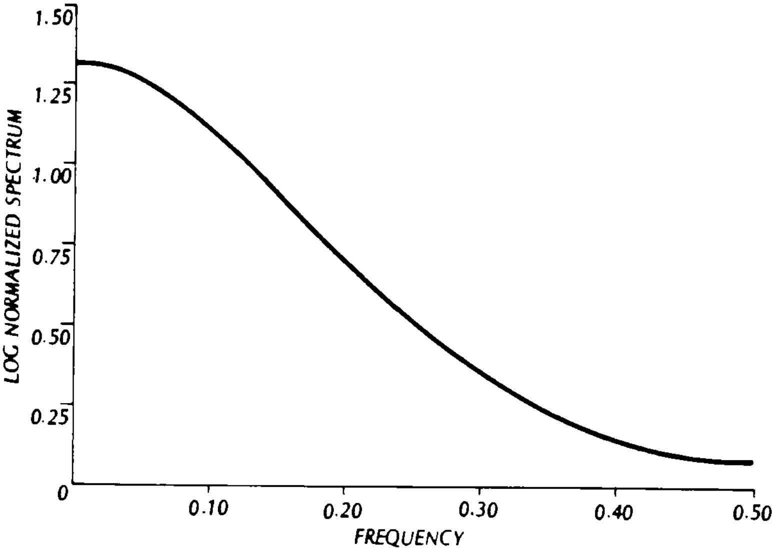


Figure 3.5.1. Log normalized spectrum for an AR(1) process with $\phi_1 = 0.3$.

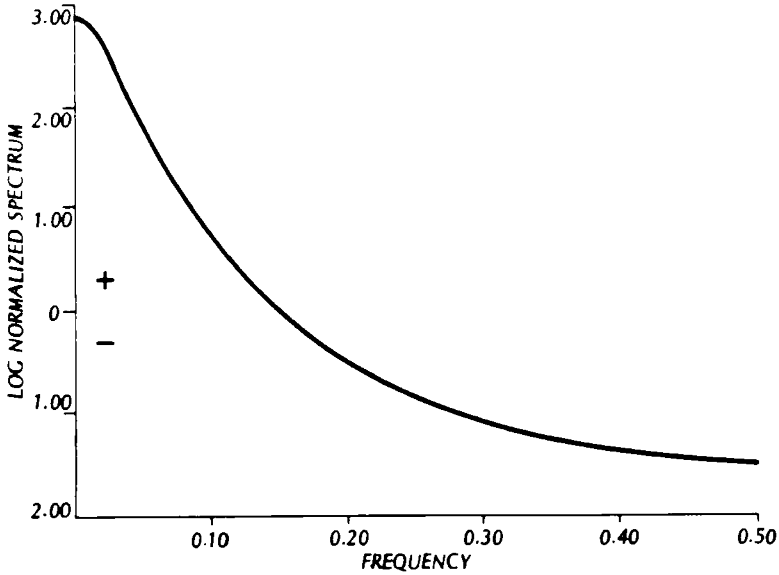


Figure 3.5.2. Log normalized spectrum for an AR(1) process with $\phi_1 = 0.8$.

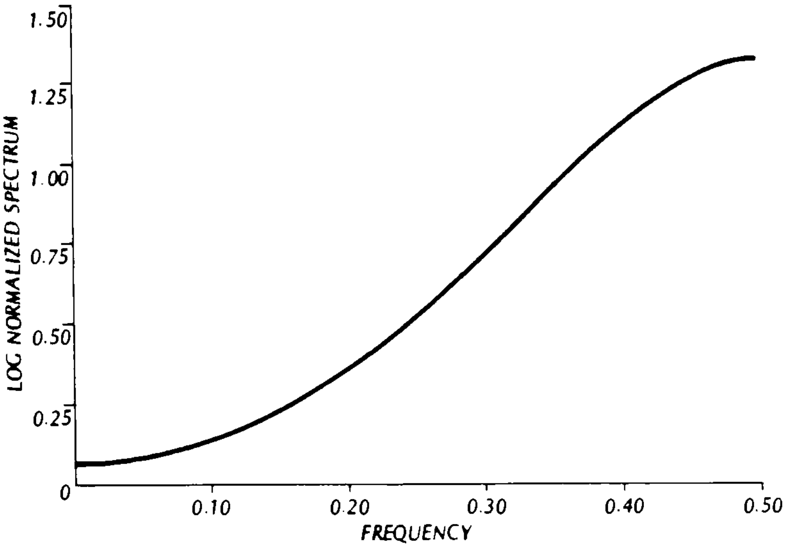


Figure 3.5.3. Log normalized spectrum for an AR(1) process with $\phi_1 = -0.3$.

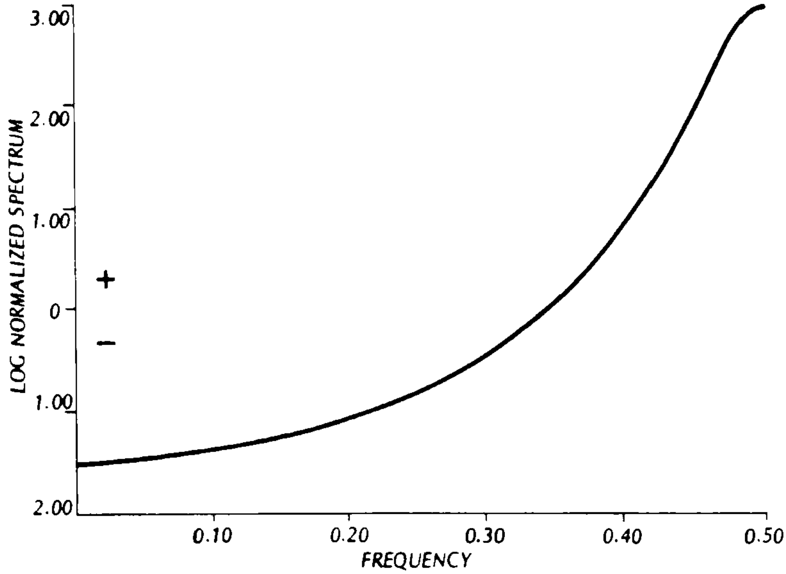


Figure 3.5.4. Log normalized spectrum for an AR(1) process with $\phi_1 = -0.8$.

physical point of view, this makes sense because the Great Lakes, that are located upstream from Ogdensburg, have a dampening effect upon extreme weather conditions that may occur in a given year. The enormous storage capacity of the Great Lakes produces a long term influence upon the flows of the St. Lawrence River and, hence, low frequency terms are dominant in the spectrum.

The MA(2) model that is fitted to the average annual temperature data in the English Midlands is given in [3.3.12]. The plot of the log normalized spectrum for this model is presented in Figure 3.5.6. The low frequency end of the spectrum is most important but the high frequency terms also assist in explaining the variability in the series. Since the fitted model is in fact close to white noise, the high points in the log normalized spectrum are spread over a wide range of frequencies.

Figure 3.5.7 is a graph of the log normalized spectrum for the ARMA(1,1) model fitted to the annual tree ring indices for Douglas Fir at the Navajo National Monument in Arizona. The difference equation for the fitted model is given in [3.4.15]. As shown in Figure 3.5.7 the spectrum is red. This could be due to the fact that the growth of a tree for a given year may be highly dependent upon the weather conditions over a long time span. For example, if the climate is favourable for healthy growth over a rather long period of time, the tree may be hardy enough to withstand severe weather patterns when they do arise without having its growth seriously retarded.

A series of 5405 tree ring widths for Bristlecone Pine at Campito Mountain in Eastern California from 3435 B.C. to 1969 A.D., is listed in units of 0.01 mm. The most appropriate ARMA model to fit to the first 500 years of this series is an ARMA(4,3) model. As can be seen for the log normalized spectrum for this model in Figure 3.5.8, there is a strong low frequency component. The peak at 0.275 cycles/year corresponds to a period of 1/0.275 = 3.6 years. When a plot of the first 500 years of the series is examined, it appears that a weak periodic component may be present in the data.

After transforming the data using a square root transformation, the most appropriate model to fit to the annual sunspot numbers is a constrained AR(9) model with ϕ_3 to ϕ_8 left out of the model. This model is given in [6.4.3]. The log normalized spectrum in Figure 3.5.9 for the sunspot model shows that the low frequencies are the most crucial for explaining the variance in the series. As noted by Granger (1957), the periodicity of the sunspot data follows a uniform distribution with a mean of about 11 years. This is confirmed by the peak in Figure 3.5.9 at a frequency of slightly less than 0.1. The cumulative periodogram for the sunspot data in Figure 2.6.3 also possesses a dramatic jump at a frequency of about 1/11.

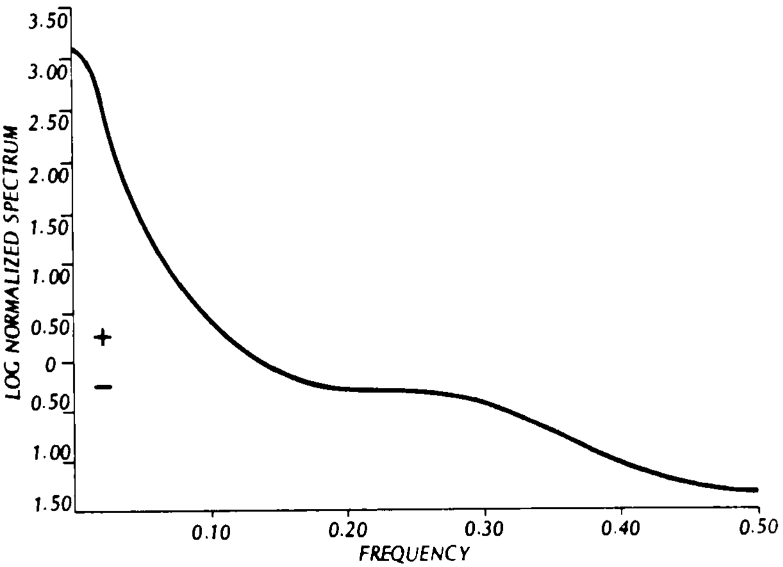


Figure 3.5.5. Log normalized spectrum for the constrained AR(3) model without ϕ_2 that is fitted to the average annual flows of the St. Lawrence River at Ogdensburg, New York.

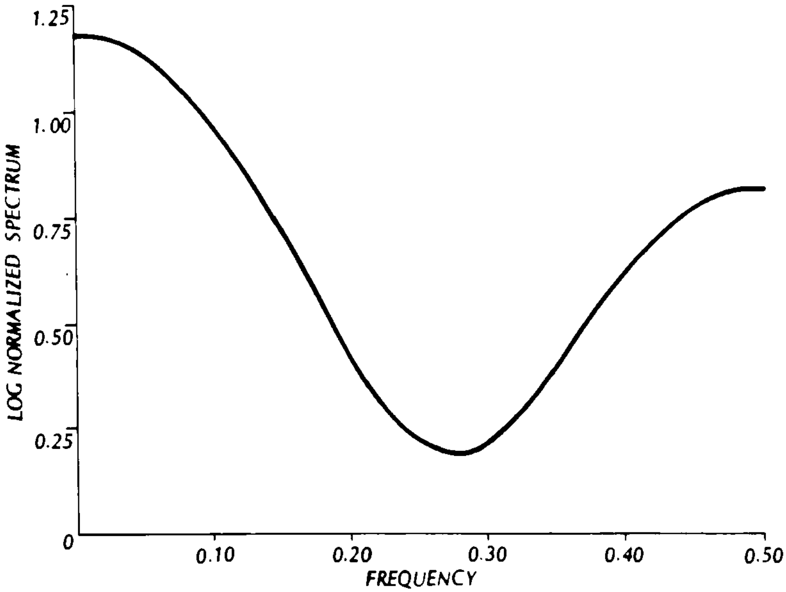


Figure 3.5.6. Log normalized spectrum for the MA(2) model fitted to the average yearly temperature data in the English Midlands.

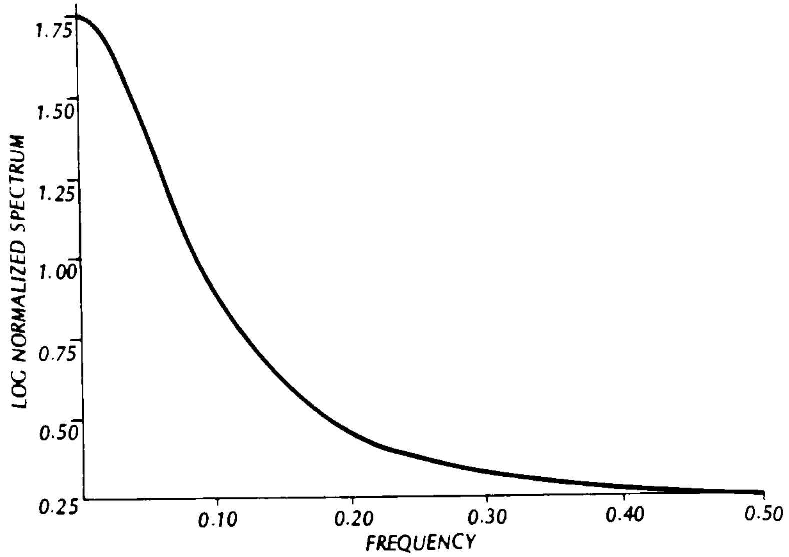


Figure 3.5.7. Log normalized spectrum for the ARMA(1,1) model fitted to the annual tree ring indices for Douglas Fir at Navajo National Monument in Arizona.

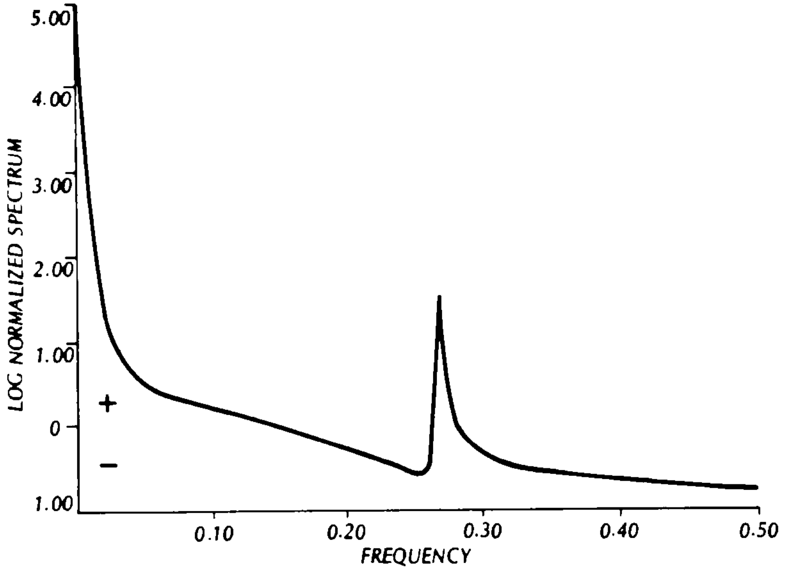


Figure 3.5.8. Log normalized spectrum for the ARMA(4,3) model fitted to the first 500 years of the Bristlecone Pine tree ring series at Campito Mountain, California.

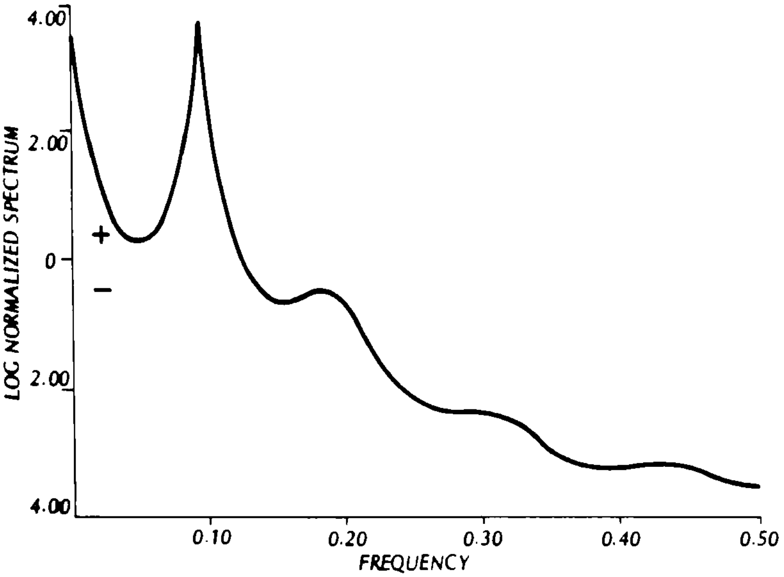


Figure 3.5.9. Log normalized spectrum for the constrained AR(9) model without ϕ_3 to ϕ_8 that is fitted to the sunspot numbers series transformed using square roots.

3.6 PHYSICAL JUSTIFICATION OF ARMA MODELS

3.6.1 Environmental Systems Model of a Watershed

The main physical components of the hydrological cycle are shown in Figure 1.4.1. As explained in Section 1.4.2, the hydrological cycle is the environmental system describing the distribution and circulation of water in all its forms on the surface of the land, underground and in the atmosphere. When modelling any part of the hydrological cycle, one would like to employ models that encapsulate the key physical characteristics of the subsystem being modelled. In other words, one would desire to use models that are physically founded and thereby properly describe the essential elements of the physical system.

For a substantial period of time, hydrologists as well as other environmental scientists have been concerned with developing a physical basis for stochastic modelling. In 1963, for example, Yevjevich examined the physical justification for using the AR(1) model in [3.2.3]. Moss and Bryson (1974) looked at the physical basis of seasonal stochastic models, which are described in Part VI of this book. Klemes (1978) as well as Salas and Smith (1981) provided a review of research on the physical foundations of stochastic models used in hydrology. Moreover, Parlange et al. (1992) explained how an AR(1) model can be formulated on the basis of the hydrologic budget and soil water transport equation, and demonstrate that the model predictions compare well with experimental results.

Fiering (1967) entertained a watershed in which the annual precipitation is decomposed into evaporation, infiltration and surface runoff. By employing the mass balance equation for the groundwater storage, he found the correlation structure of annual streamflow as a function of the correlation structure of precipitation which was assumed to be independent or else AR(1). Salas and Smith (1981) demonstrated that the conceptual watershed model of Thomas and Fiering leads to ARMA streamflows and ARMA groundwater storage. The objective of this section is to point out some of the main findings of Salas and Smith (1981) so that the reader can fully appreciate the *physical justification* for employing ARMA models in hydrology.

Figure 3.6.1 displays the environmental systems model for a *watershed* that Salas and Smith (1981) utilize in their research. This systems model is, of course, a component of the overall hydrological cycle depicted in Figure 1.4.1. In essence, the physical systems model in Figure 3.6.1 shows how precipitation is transformed into runoff or annual riverflow.

Following the notation provided by Salas and Smith (1981) for the environmental model of the watershed shown in Figure 3.6.1, let x_t represent the precipitation in year t. Assume that an amount bx_t of the precipitation evaporates and an amount ax_t infiltrates through the soil into groundwater storage. Therefore, $(1 - a - b)x_t = dx_t$ represents the surface runoff that flows into the rivers and streams. Moreover, let S_{t-1} be the groundwater storage at the start of year t and assume that cS_{t-1} is the groundwater contribution to runoff. In the above algebraic description of the watershed model, it is necessary that $0 \leq a,b,c,d \leq 1$ and $0 \leq a + b \leq 1$.

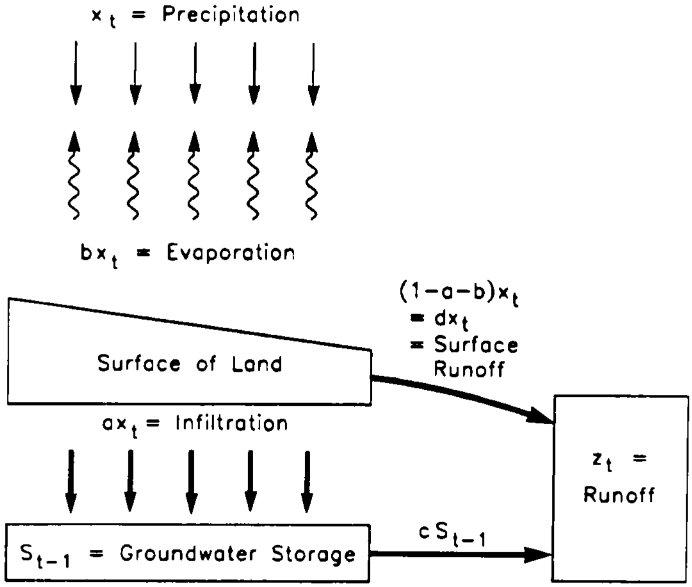


Figure 3.6.1. Environmental systems model of a watershed.

As shown in Figure 3.6.1, the total runoff or riverflow z_t is composed of the direct surface runoff dx_t plus the groundwater contribution cS_{t-1}. Accordingly,

$$z_t = cS_{t-1} + dx_t \qquad [3.6.1]$$

Furthermore, the *mass balance equation* for the groundwater storage is

$$S_t = (1-c)S_{t-1} + ax_t \qquad [3.6.2]$$

The above two equations can be combined to obtain (Salas and Smith, 1981)

$$z_t = (1-c)z_{t-1} + dx_t - [d(1-c) - ac]x_{t-1} \qquad [3.6.3]$$

When writing down the difference equations for the AR, MA, and ARMA models in Sections 3.2 to 3.4, respectively, the mean level μ is subtracted from the variable z_t being modelled. Because of this, the theoretical mean of the a_t innovations in these models is zero. In order to compare the results of this section to the ARMA models, it is convenient to write equations [3.6.1] to [3.6.3] in a similar fashion. More specifically, let μ, μ_x and μ_s be the theoretical means for the variables z_t, x_t and S_t, respectively. By replacing z_t, x_t and S_t by $(z_t - \mu)$, $(x_t - \mu_x)$ and $(S_t - \mu_s)$, respectively, equations [3.6.1] to [3.6.3] can be equivalently rewritten as

$$z_t - \mu = c(S_{t-1} - \mu_s) + d(x_t - \mu_x) \qquad [3.6.4]$$

$$S_t - \mu_s = (1-c)(S_{t-1} - \mu_s) + a(x_t - \mu_x) \qquad [3.6.5]$$

$$z_t - \mu = (1-c)(z_{t-1} - \mu) + d(x_t - \mu_x) - [d(1-c) - ac](x_{t-1} - \mu_x) \qquad [3.6.6]$$

Based upon three different models for the precipitation, Salas and Smith (1981) derive the models for the corresponding groundwater storage and riverflows. Below, the results of their research are summarized for the three cases of independent, AR(1) and ARMA(1,1) precipitation.

3.6.2 Independent Precipitation

If the precipitation is independent, it can be written as

$$(x_t - \mu_x) = a_t \qquad [3.6.7]$$

where μ_x is the mean of the total amount of precipitation x_t falling in year t and a_t is IID$(0,\sigma_a^2)$ as in [3.2.2]. Substituting $(x_t - \mu_x)$ from [3.6.7] into [3.6.5] produces

$$S_t - \mu_s = (1-c)(S_{t-1} - \mu_s) + a(a_t) \qquad [3.6.8]$$

In terms of the groundwater storage variable S_t, the above relationship is simply an AR(1) model. When the AR(1) model in [3.6.8] is compared to the one in [3.2.1] notice that S_t replaces z_t, $(1-c) = \phi_1$ and a constant a instead of unity is in front of the innovation term.

To find the relationship for riverflow, replace $(x_t - \mu_x)$ by a_t in [3.6.6] to obtain

$$z_t - \mu = (1-c)(z_{t-1} - \mu) + da_t - [d(1-c) - ac]a_{t-1} \qquad [3.6.9]$$

The above expression for total yearly flow z_t is simply an ARMA(1,1) model defined in [3.4.1]. One can employ [3.4.17] to write the theoretical ACF for the ARMA(1,1) model as

$$\rho_k = (1-c)\rho_{k-1} \text{ for } k > 1 \qquad [3.6.10]$$

where ρ_k is the theoretical ACF at lag k.

In summary, independent precipitation produces AR(1) storage as shown in [3.6.8]. Additionally, this kind of precipitation causes the ARMA(1,1) flow given in [3.6.9].

3.6.3 AR(1) Precipitation

As is demonstrated below for the watershed model in Figure 3.6.1, AR(1) precipitation causes AR(2) groundwater storage and ARMA(2,1) runoff. From [3.2.1], an AR(1) model for the precipitation x_t is written as

$$(x_t - \mu_x) = \phi_1(x_{t-1} - \mu_x) + a_t \qquad [3.6.11]$$

To determine the type of groundwater storage that this precipitation creates, substitute [3.6.11] into [3.6.5] to get

$$S_t - \mu_s = (1 - c + \phi_1)(S_{t-1} - \mu_s) - (1-c)\phi_1(S_{t-2} - \mu_s) + (a)a_t \qquad [3.6.12]$$

In terms of storage, [3.6.12] is an ARMA(1,1) model.

By combining [3.6.11] and [3.6.6], the riverflow generated by AR(1) precipitation is

$$z_t - \mu = (1 - c + \phi_1)(z_{t-1} - \mu) - (1-c)\phi_1(z_{t-2} - \mu)$$

$$+ (d)a_t - [d(1-c) - ac]a_{t-1} \qquad [3.6.13]$$

Hence, AR(1) precipitation causes ARMA(2,1) riverflows. From [3.4.13], the theoretical ACF for this ARMA(2,1) model is

$$\rho_k = (1 - c + \phi_1)\rho_{k-1} - (1-c)\phi_1\rho_{k-2} \quad \text{for } k > 1 \qquad [3.6.14]$$

3.6.4 ARMA(1,1) Precipitation

The ARMA(1,1) model for the precipitation x_t in Figure 3.6.1 is written as

$$(x_t - \mu_x) = \phi_1(x_{t-1} - \mu_x) + a_t - \theta_1 a_{t-1} \qquad [3.6.15]$$

By substituting [3.6.15] into [3.6.5], the resulting groundwater storage is found to be

$$S_t - \mu_s = (1 - c + a\phi_1)(S_{t-1} - \mu_s) + (1-c)a\phi_1(S_{t-2} - \mu_s) + (a)a_t - a\theta_1 a_{t-1} \qquad [3.6.16]$$

which is an ARMA(2,1) model. When [3.6.15] is combined with [3.6.6], the model for riverflow is

$$\begin{aligned} z_t - \mu &= (1 - c + \phi_1)(z_{t-1} - \mu) - (1 - c)\phi_1(z_{t-2} - \mu) \\ &\quad + da_t - [d(1 - c + \theta_1) - ac]a_{t-1} - \theta_1[ac - d(1 - c)]a_{t-2} \end{aligned} \qquad [3.6.17]$$

which corresponds to an ARMA(2,2) model. The theoretical ACF for the model in [3.6.17] is obtained from [3.4.13] as

$$\rho_k = (1 - c + \phi_1)\rho_{k-1} - (1 - c)\phi_1\rho_{k-2} \text{ for } k > 2 \qquad [3.6.18]$$

Table 3.6.1 summarizes the kinds of groundwater storage and streamflow models that are created by the three different types of precipitation investigated in Sections 3.6.2 to 3.6.4. As noted earlier, these results were originally derived by Salas and Smith (1981) for the environmental systems model of the watershed displayed in Figure 3.6.1. The findings clearly demonstrate that ARMA models possess a valid physical basis for modelling this kind of hydrologic system. Consequently, in Parts III, IV and V of the book, ARMA models are fitted directly to annual riverflow and other types of yearly environmental time series.

Table 3.6.1. Physical basis of ARMA models in hydrology.

Types of Models for Precipitation	Resulting Models	
	Groundwater Storage	Streamflow Runoff
Independent	AR(1)	ARMA(1,1)
AR(1)	AR(2)	ARMA(2,1)
ARMA(1,1)	ARMA(2,1)	ARMA(2,2)

3.7 CONCLUSIONS

The AR and MA classes of models of Sections 3.2 and 3.3, respectively, are members of the general family of ARMA models defined in Section 3.4. These models possess sound theoretical designs and their important theoretical properties are known. For example, the theoretical ACF's for AR, MA, and ARMA models are derived in this chapter and a simple algorithm for calculating the theoretical ACF of any ARMA model is given in Appendix A3.2. Knowledge of the theoretical ACF structure of ARMA models is required for identifying the most appropriate type of ARMA model to fit to a given data set. As explained in Part III, well developed model construction tools are available for fitting ARMA models to stationary nonseasonal time series by following the identification, estimation and diagnostic check stages of model building. Practical applications in Part III clearly demonstrate that ARMA models are ideally suited for describing stationary annual riverflow series as well as other kinds of environmental data sets.

In addition to having a rigorous theoretical design and possessing comprehensive model building tools, ARMA models possess other inherent assets for ensuring their successful application in the environmental sciences. Firstly, the results of Section 3.6 confirm that there is valid physical justification for employing ARMA models for fitting to yearly hydrologic time series. For example, from Table 3.6.1 one can see that if the annual precipitation is ARMA(1,1), then the groundwater storage must be ARMA(2,1) and the yearly streamflow runoff is ARMA(2,2).

Secondly, in Chapter 10 it is clearly demonstrated using annual hydrologic data and simulation experiments that ARMA models provide a logical explanation for the famous Hurst phenomenon. More specifically, ARMA models are shown to preserve statistically what are called the Hurst statistics, which are statistics that reflect the long term storage capacity of reservoirs. Thirdly, forecasting experiments using yearly hydrologic and other kinds of time series in Chapter 8, show that ARMA models forecast at least as well and usually better than their competitors. Finally, the basic ARMA model of Chapter 3 provides the solid foundations for developing the long memory, seasonal, transfer function-noise, intervention and multivariate models of Chapter 10, and Parts VI to IX, respectively. In fact, by introducing what is called the differencing operator to remove nonstationarity, the ARMA model is extended in Chapter 4 so that it can handle nonstationary annual time series.

APPENDIX A3.1

ALGORITHM FOR ESTIMATING THE PARTIAL AUTOCORRELATION FUNCTION

The *Pagano algorithm* (1972) uses the following steps to estimate the PACF up to lag p for a specified time series.

1. Determine the modified Cholesky decomposition (Wilkinson, 1965, p. 229) of the estimated autocorrelation matrix $\mathbf{R}_p$ given by

$$\mathbf{R}_p = \begin{bmatrix} 1 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & 1 & r_1 & \cdots & r_{p-2} \\ r_2 & r_1 & 1 & \cdots & r_{p-3} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & 1 \end{bmatrix} \qquad [A3.1.1]$$

where r_k is estimated using [2.5.9] and the theoretical ACF is defined in [2.5.4]. The modified Cholesky decomposition of $\mathbf{R}_p$ is

$$\mathbf{R}_p = \mathbf{L}_p \mathbf{D}_p \mathbf{L}_p^T \qquad [A3.1.2]$$

where $\mathbf{L}_p$ is a unit lower triangular matrix defined by

$$\mathbf{L}_p = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ l_{p1} & l_{p2} & l_{p3} & \cdots & 1 \end{bmatrix},$$

$\mathbf{D}_p$ is a diagonal matrix where d_{kk} is the kth typical diagonal entry which is obtained from

$$d_{ii} l_{ki} = r_{k-i} - \sum_{j=1}^{i} b_{kj} l_{jj} = b_{ki}\ , \quad i = 1,2,\ldots,k-1$$

$$d_{kk} = 1 - \sum_{j=1}^{k-1} b_{kj} l_{kj}$$

and

$$d_{11} = 1$$

where the sequence b_{kj} is defined by the algorithm.

2. Solve the triangular system of equations given by

$$\mathbf{L}_p \alpha_p = \mathbf{r}_p \qquad [A3.1.3]$$

where the unknown vector is

$$\alpha_p^T = (\alpha_1, \alpha_2, \ldots, \alpha_p)$$

and

$$\mathbf{r}_p^T = (r_1, r_2, \ldots, r_p)$$

3. Calculate the estimates $\hat{\phi}_{kk}$ of the PACF using

$$\hat{\phi}_{kk} = \frac{\alpha_k}{d_{kk}}\ , \quad k = 1,2,\ldots,p \qquad [A3.1.4]$$

4. If the $\hat{\phi}_{kj}(j = 1,2,\ldots,k)$ are required for some $k \le p$, they can be determined by solving the triangular system of equations

$$\mathbf{L}_k^T = \begin{pmatrix} \hat{\phi}_{k1} \\ \hat{\phi}_{k2} \\ \cdot \\ \cdot \\ \cdot \\ \hat{\phi}_{kk} \end{pmatrix} = \begin{pmatrix} \hat{\phi}_{11} \\ \hat{\phi}_{22} \\ \cdot \\ \cdot \\ \cdot \\ \hat{\phi}_{kk} \end{pmatrix} \qquad [A3.1.5]$$

From [3.2.13], the estimate for the variance of the white noise sequence for an AR model of order k is given by

$$\hat{\sigma}_a^2(k) = c_0 - \hat{\phi}_{k1}c_1 - \hat{\phi}_{k2}c_2 - \cdots - \hat{\phi}_{kk}c_k \qquad [A3.1.6]$$

Alternatively, the white noise variance for an AR(k) model may be estimated recursively by employing

$$\sigma_a^2(k) = \sigma_a^2(k-1)(1-\phi_{kk}^2) \qquad [A3.1.7]$$

where

$$\sigma_a^2(0) = c_0$$

which is the sample variance calculated using [2.5.2] for the given series.

APPENDIX A3.2

THEORETICAL ACF FOR AN ARMA PROCESS

When the parameters of either a nonseasonal or seasonal ARMA process are known, the following algorithm of McLeod (1975) can be employed to determine the theoretical autocovariance, γ_k, and also the theoretical ACF, ρ_k. For the case of a nonseasonal ARMA(p,q) model, the algorithm is as follows:

1. Set $r = max(p,q)$ and $\phi_0 = \theta_0 = -1$, $c_0 = 1$.

2. Then calculate

$$c_k = -\theta_k + \sum_{i=1}^{min(p,k)} \phi_i c_{k-i} \text{ for } k = 1,2,\ldots,q \ .$$

3. Set $b_k = -\sum_{i=k}^{q} \theta_i c_{i-k}$ for $k = 0,1,\ldots,q$, and set $b_k = 0$ if $k > q$.

4. If $p = 0$ then set $\gamma_k = b_k\sigma_a^2$ for $k = 0,1,\ldots,q$; otherwise

$$\begin{pmatrix}\gamma_0\\ \gamma_1\\ \cdot\\ \cdot\\ \cdot\\ \gamma_r\end{pmatrix} = -\mathbf{A}^{-1}\begin{pmatrix}b_0\\ b_1\\ \cdot\\ \cdot\\ \cdot\\ b_r\end{pmatrix}\sigma_a^2$$

where A is the $(r+1)$ by $(r+1)$ matrix with (i,j) entry:

$$A_{ij} = \begin{cases}\phi_{i-j},\ j = 1,\ i = 1,2,\ldots,r+1\\ \phi_{i-j} + \phi_{i+j-2},\ j = 2,3,\ldots,r+1,\ i = 1,2,\ldots,r+1\end{cases}$$

$$\phi_k = \begin{cases} \phi_k, k = 0,1,\ldots,p \\ 0, \textit{ otherwise} \end{cases}$$

5. For $k > r = max(p,q)$ calculate γ_k recursively from

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p}$$

6. Divide γ_k by γ_0 to obtain ρ_k for $k = 1,2,\ldots$.

PROBLEMS

3.1 Two stochastic processes, z_{1t} and z_{2t}, have theoretical autocovariance functions at lag k of γ_{1k} and γ_{2k}, respectively, where $\gamma_{1k} = 0$ for $k > 2$ and $\gamma_{2k} = 0$ for $k > 3$. Derive the theoretical autocorrelation function (ACF) for the process $z_{3t} = z_{1t} + bz_{2t}$ in terms of the theoretical autocovariance functions for z_{1t} and z_{2t} where b is a constant. Assume that z_{1t} and z_{2t} are independent of one another.

3.2 Using equations, clearly explain how Cholesky decomposition works.

3.3 For an AR(2) process given by

$$(1 - 1.1B + 0.24B^2)z_t = a_t$$

a) calculate ρ_1,

b) using ρ_1 as starting values for the difference equation in [3.2.9], determine ρ_k, $k = 2,3,\ldots,12$,

c) plot the ACF for this model and comment about its behaviour.

3.4 A constrained AR(3) model without the second AR parameter, ϕ_2, is written as

$$(1 - \phi_1 B - \phi_3 B^3)(z_t - \mu) = a_t$$

From basic principles, derive the Yule-Walker equations for this specific AR model.

3.5 Compare the advantages and disadvantages of using the following three methods for estimating the PACF. Briefly explain how each method works.

a) Cramer's rule,

b) Durbin's method [see Box and Jenkins (1976) and also Durbin (1960)],

c) Pagano's (1972) technique.

3.6 Using equations, explain how the Burg algorithm works for estimating the parameters of an AR(p) model. As an example, show how the Burg algorithm is employed for estimating the parameters of an AR(2) model.

3.7 From first principles, derive the theoretical ACF for a MA(2) process. Using the Yule-Walker equations, determine the theoretical PACF for this process.

3.8 For the ARMA(1,1) process in [3.4.2], derive the two main equations that are required to solve for γ_k, the theoretical autocovariance function of this process. Use these equations to solve for γ_k, $k = 0,1,2,\ldots$.

3.9 An ARMA model is written as

$$(1 - 0.8B + 0.12B^2)z_t = (1 - 0.2B)a_t$$

Prove whether or not this model is stationary.

3.10 Using the hints given with [3.4.20], prove that for stationarity, the roots of $\phi(B) = 0$ must lie outside the unit circle.

3.11 An ARMA(p,q) model is given as

$$(1 - 0.7B)z_t = (1 - 0.4B - 0.21B^2)a_t$$

Prove whether or not this model is invertible.

3.12 The constrained AR(3) model for the annual flows of the St. Lawrence River at Ogdensburg, New York, is given in [3.2.19] as

$$(1 - 0.619B - 0.177B^3)(z_t - 6818.63) = a_t$$

Write this model in inverted form.

3.13 For the ARMA(1,1) model in [3.4.2], determine

a) the parameters ψ, ψ_2 and ψ_3 in the random shock operator, and

b) the parameters π_1, π_2 and π_3 in the inverted operator.

3.14 Express the model given by

$$(1 - 0.6B)(z_t - 15) = (1 - 0.8B)a_t$$

in

a) random shock form, and

b) inverted form.

3.15 Prove that the Box-Cox power transformation in [3.4.30] is continuous at $\lambda = 0$.

3.16 One method for causing non-normal data to become normal is to invoke the Box-Cox transformation in [3.4.30]. Subsequent to this, an ARMA(p,q) model can be fitted to the data that now approximately follow a normal distribution. Other approaches are also available for modelling non-normal data. Describe other transformations suggested by Jain and Singh (1986) for applying to non-normal data sets. Briefly explain how Lewis (1985) and other authors cited in his paper handle the problem of modelling data that do not follow a normal distribution.

3.17 By employing [3.5.7] and [3.5.10], obtain the autocovariance function and normalized spectrum for an

a) AR(1) model, and

b) ARMA(1,1) model.

3.18 An environmental systems model of a watershed is depicted in Figure 3.6.1. Suppose that the precipitation input to this system is ARMA(2,1). Derive the types of models that this precipitation causes for groundwater storage and streamflow runoff. Write down the theoretical ACF's for the precipitation, groundwater storage and runoff models.

3.19 Section 3.6 explains how ARMA models can realistically describe the watershed system displayed in Figure 3.6.1. Investigate the validity of ARMA models for describing another environmental system such as a system of reservoirs or a sewage treatment facility.

REFERENCES

DATA SETS

Stokes, M. A., Drew, L. G. and Stockton, C. W. (1973). Chronology Series 1, Laboratory of Tree Ring Research, University of Arizona, Tucson, Arizona.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960.* Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology paper No. 1, Colorado State University, Fort Collins, Colorado.

DATA TRANSFORMATIONS

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 26:211-252.

Jain, D. and Singh, V.P. (1986). A comparison of transformation methods for flood frequency analysis. *Water Resources Bulletin* 22(6):903-912.

PHYSICAL BASIS OF TIME SERIES MODELS

Klemes, V. (1978). Physically based stochastic hydrologic analysis. In *Advances in Hydroscience* 11:285-356, Academic Press, New York.

Moss, M. E. and Bryson, M. C. (1974). Autocorrelation structure of monthly streamflows. *Water Resources Research,* 10:737-744.

Parlange, M. B., Katul, G. G., Cuenca, R. H., Kavvas, M. L., Nielsen, D. R. and Mata, M. (1992). Physical basis for a time series model of soil water content. *Water Resources Research,* 28(9):2437-2446.

Salas, J. D. and Smith, R. A. (1981). Physical basis of stochastic models of annual flows. *Water Resources Research,* 17(2):428-430.

STOCHASTIC HYDROLOGY

Fiering, M. B. (1967). *Streamflow Synthesis*. Harvard University Press, Cambridge, Massachusetts.

Hipel, K.W. and McLeod, A.I. (1978). Preservation of the rescaled adjusted range, 2, Simulation studies using Box-Jenkins models. *Water Resources Research* 14(3):509-516.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770-808.

Hurst, H. E. (1956). Methods of using long-term storage in reservoirs. *Proceedings of the Institute of Civil Engineers*, 1:519-543.

McLeod, A. I. and Hipel, K. W. (1978). Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon. *Water Resources Research*, 14(3):491-508.

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, applications. *Water Resources Research*, 13(3):577-586.

TIME SERIES ANALYSIS

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716-723.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.

Barndorff-Nielsen, O. and Schou, G. (1973). On the parameterization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis* 3:408-419.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Mass.

Burg, J. P. (1975). *Maximum Entropy Spectral Analysis*. Ph.D. dissertation, Department of Geophysics, Stanford University, Stanford, California.

Durbin, J. (1960). The fitting of time series models. *Revue de L'Institut International de Statistique*, 28(3):233-244.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. John Wiley, New York.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysics Journal*, 126:152-158.

Haggan, V. and Oyetunji, O. B. (1984). On the selection of subset autoregressive time series models. *Journal of Time Series Analysis*, 5(2):103-113.

Haykin, S. (1990). *Modern Filters*. MacMillan, New York.

Healy, M. J. R. (1968). Algorithm AS6, triangular decomposition of a symmetric matrix. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 17:195-197.

Hipel, K.W., McLeod, A.I. and Lennox, W.C. (1977). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research* 13(3):567-575.

Kleiner, B., Martin, R. D. and Thomson, D. J. (1979). Robust estimation of power spectrum. *Journal of the Royal Statistical Society*, 41(3):313-351.

Lewis, P.A.W. (1985). Some simple models for continuous variate time series. *Water Resources Bulletin* 21(4):635-644.

McLeod, A. I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 24(2):255-256.

Otnes, R. K. and Enochson, L. (1972). *Digital Time Series Analysis*. Wiley.

Pagano, M. (1972). An algorithm for fitting autoregressive schemes. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 21:274-281.

Pagano, M. (1973). When is an autoregressive scheme stationary. *Communications in Statistics*, 1(6):533-544.

Pandit, S.M. and Wu, S.M. (1983). *Time Series and System Analysis with Applications*. Wiley, New York.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:68-84.

Robinson, E. A. (1967). *Multichannel Time Series Analysis with Digital Computer Programs*. Holden-Day, San Francisco.

Schur, I. (1917). Uber potenzreihen, die in innem des einheitskreises beschrankt sind. *J. Reine Agnew. Math.*, 147:205-232.

Siddiqui, M. M. (1958). On the inversion of the sample covariance matrix in a stationary autoregressive process. *Annals of Mathematical Statistics*, 29:585-588.

Walker, G. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society, A*, 131:518-532.

Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

Yu, G. H. and Lin, Y. C. (1991). A methodology for selecting subset autoregressive time series models. *Journal of Time Series Analysis*, 12(4):363-373.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer sunspot numbers. *Phil. Transactions of the Royal Society, Series A*, 226:267-298.

CHAPTER 4

NONSTATIONARY NONSEASONAL MODELS

4.1 INTRODUCTION

When considering annual hydrological and other natural time series of moderate lengths (perhaps a few hundred years), it is often reasonable to assume that a stationary model can adequately model the data. For example, in Section 10.6.2, stationary ARMA models are fitted to 23 time series which are measured from six different types of natural phenomena that vary in length from 96 to 1164 years. The ability to detect statistical characteristics of a time series which change significantly over time may only become possible when the yearly records cover a sufficiently long time horizon. For example, if past climatic records were available or could be constructed for a given location in North America, the results would probably support the hypothesis of climatic nonstationarity over a long time span. Certainly, as the ice sheets advanced and retreated over the North American continent during the past one hundred thousand years, average annual temperatures and other climatic factors changed significantly over time.

Some types of annual time series which are studied in water resources engineering, could be *nonstationary* even over a short time interval. For instance, the average annual cost of hydroelectric power and the annual consumption of water of an expanding metropolis constitute two time series which increase in magnitude over time. In general, time series that reflect the socio - economic aspects of water resources planning may be nonstationary over any time interval being considered.

When modelling nonstationary data, a common procedure is to first remove the nonstationarity by using a suitable technique. Following this, a stationary model can be fit to the resulting stationary time series. This general approach is used in this chapter for nonseasonal models and also in Chapter 12 for a certain class of seasonal models.

4.2 EXPLOSIVE NONSTATIONARITY

If an ARMA(p,q) process is *stationary*, all of the roots of the characteristic equation $\phi(B)=0$ must lie outside the unit circle (see Section 3.2.2). Consequently, when a process is *nonstationary* at least one of the roots of $\phi(B)=0$ must lie on or within the unit circle. If at least one root is inside the unit circle, the process is said to possess *explosive nonstationarity*. When none of the roots are within the unit circle but at least one of the roots lies on the unit circle, this is referred to as *homogeneous nonstationarity*.

For the case of an ARMA(1,1) process in [3.4.1], it is necessary that the root ϕ_1^{-1} of $(1-\phi_1 B)=0$ possess an absolute magnitude which is greater than unity or, equivalently, $|\phi_1|<1$ in order to have stationarity. On the other hand, when a process with one AR and one MA parameter is nonstationary, the root ϕ_1^{-1} must lie either on or inside the unit circle and hence $|\phi_1|\geq 1$. Suppose, for example, a model is given as

$$z_t - \phi_1 z_{t-1} = a_t - 0.70a_{t-1}$$

or, equivalently,

$$(1 - \phi_1 B)z_t = (1 - 0.70B)a_t \qquad [4.2.1]$$

where a_t is normally independently distributed with a mean of zero and a variance of one [i.e., NID(0,1)]. If $\phi_1 = 1.1$, the root of $1 - 1.1B = 0$ is $1/1.1$ and hence the process possesses explosive nonstationarity. When z_1 is assigned a value of, say, 100, z_t can be simulated using

$$z_2 - 1.1z_1 = a_2 - 0.70a_1 \qquad [4.2.2]$$

where the a_t's are randomly generated on a computer (see Section 9.2). By substituting $t = 3,4,\ldots,20$, into [4.2.1], a sequence of 20 synthetic data points can be obtained where $z_1 = 100$. A plot of 20 simulated values for z_t is shown in Figure 4.2.1. Notice how the series increases greatly over time due to the fact that the root of the characteristic equation lies just inside the unit circle. If ϕ_1 is given a value of 1.5, a simulated series can be even more explosive than that presented in Figure 4.2.1. The simulated sequence of 20 values in Figure 4.2.2 was obtained using [4.2.1] with $\phi_1 = 1.5$ and a starting value of $z_1 = 100$. In that figure, the series increases exponentially with time and the last synthetic data point has a magnitude which is close to 24,000.

4.3 HOMOGENEOUS NONSTATIONARITY

The ARIMA (autoregressive integrated moving average) model is defined in the next subsection for modelling an annual time series possessing homogeneous nonstationarity. As explained in Section 4.3.2, the theoretical ACF for an ARIMA model containing nonstationarity dies off slowly. Consequently, if the sample ACF of a given annual time series attenuates, this may indicate the presence of nonstationarity and the need to fit an ARIMA model to the series. Three kinds of time series are employed in Section 4.3.3 to demonstrate how the sample ACF dies off slowly for a nonstationary series and how to fit an ARIMA model to each series. Finally, Section 4.3.4 describes three equivalent formulations of the ARIMA model.

4.3.1 Autoregressive Integrated Moving Average Model

When at least one of the roots of the characteristic equation lies on the unit circle but none of the roots are inside the unit circle, this produces a milder type of nonstationarity than the explosive case. This is referred to as *homogeneous nonstationarity* because, except for a local level and slope, often portions of a simulated series will be similar to other sections. For example, when ϕ_1 is set equal to unity in [4.2.1] the model becomes

$$z_t - z_{t-1} = a_t - 0.70a_{t-1}$$

or equivalently

$$(1-B)z_t = (1 - 0.70B)a_t \qquad [4.3.1]$$

where a_t is NID(0,1). Notice that the single root of $(1 - B) = 0$ is of course unity and hence the model possesses homogeneous nonstationarity. By choosing a starting value of $z_1 = 100$ and

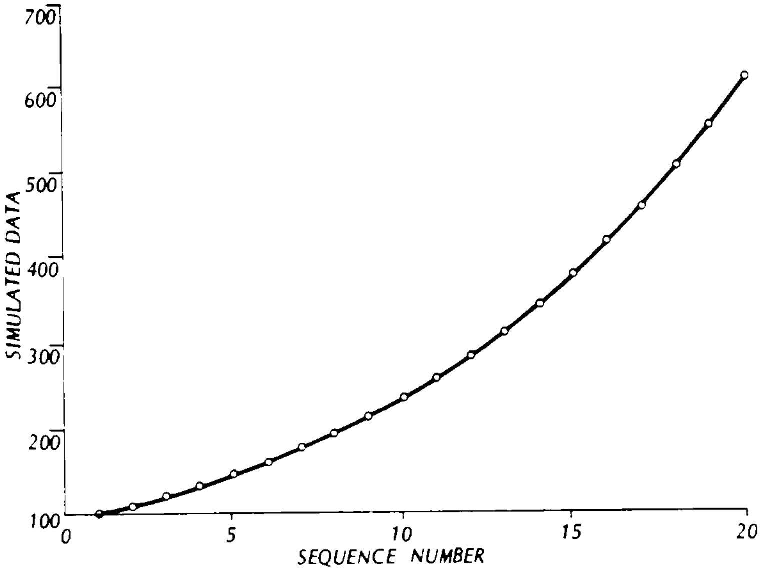


Figure 4.2.1. Simulated data for the model in [4.2.1] with $\phi_1 = 1.1$ and $z_1 = 100$.

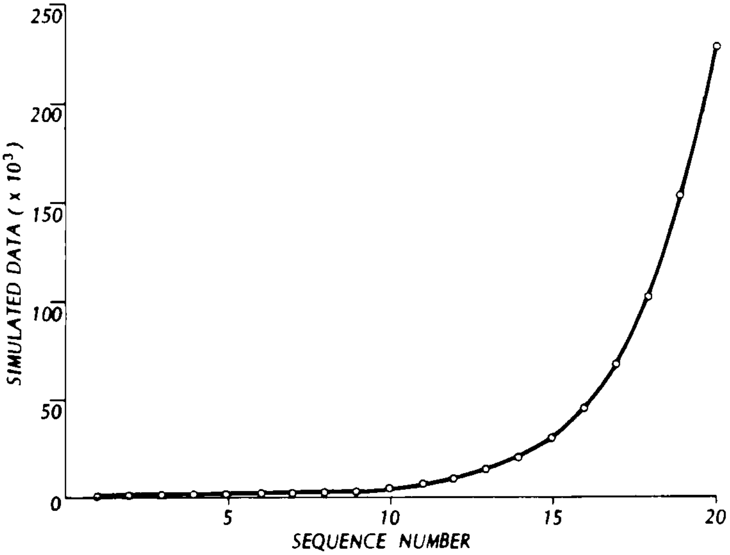


Figure 4.2.2. Simulated data for the model in [4.2.1] with $\phi_1 = 1.5$ and $z_1 = 100$.

having the computer generate the a_t's, a sequence of 20 simulated values can be obtained as shown in Figure 4.3.1. It can be seen that this realization behaves in a much more restrained fashion than those shown in Figures 4.2.1 and 4.2.2. This kind of behaviour is typical of many types of socio - economic series which are encountered in practical applications and therefore the modelling of homogeneous nonstationarity has received widespread attention (Box and Jenkins, 1976).

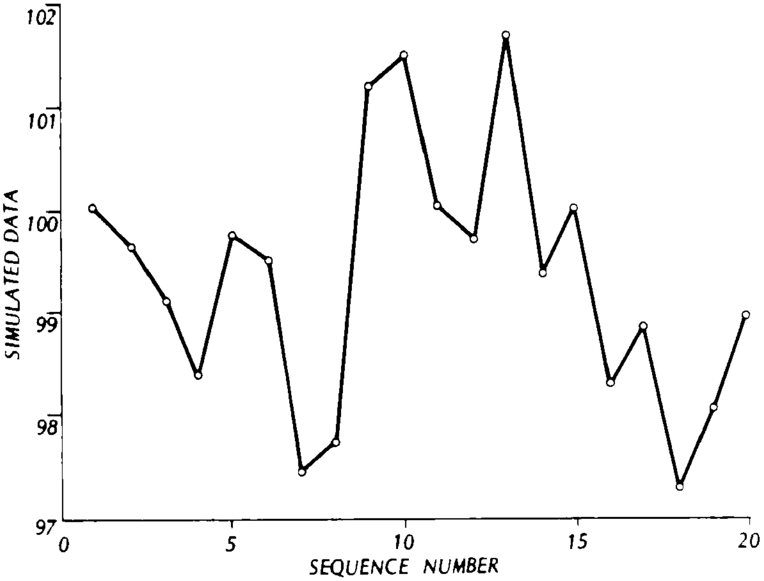


Figure 4.3.1. Simulated data for the model in [4.3.1] that possesses homogeneous nonstationarity.

The operator $\nabla = (1 - B)$ in [4.3.1] is referred to as the *differencing operator* because the root of $(1 - B) = 0$ lies on the unit circle. When ∇ operates on $(z_t - \mu)$ the level μ disappears due to the nonstationarity as is shown by

$$(1 - B)(z_t - \mu) = (z_t - \mu) - (z_{t-1} - \mu) = z_t - z_{t-1} = (1 - B)z_t \qquad [4.3.2]$$

When a time series of length N is differenced using [4.3.2], adjacent time series values are subtracted from each other to obtain a sequence of length $N - 1$. This differencing procedure can be repeated just enough times to produce a stationary series labelled w_t. In general, a time series may be differenced d times to produce a stationary series of length $n = N - d$ given by

$$w_t = (1 - B)^d z_t = \nabla^d z_t$$

If the original z_t time series is transformed by a Box-Cox transformation as explained in Section 3.4.5, the stationary w_t series is formed by differencing the transformed series and is calculated using

$$w_t = (1-B)^d z_t^{(\lambda)} \qquad [4.3.3]$$

When homogeneous nonstationarity is present, it is reasonable to assume that the w_t series in [4.3.3] can be modelled by the stationary ARMA(p,q) model in [3.4.3] such that

$$\phi(B)w_t = \theta(B)a_t \qquad [4.3.4]$$

where the roots of $\phi(B)=0$ lie outside the unit circle for stationarity of the w_t sequence, the d roots of $(1-B)^d$ are on the unit circle due to the homogeneous nonstationarity of the $z_t^{(\lambda)}$ series in [4.3.3], and the roots of $\theta(B)=0$ lie outside the unit circle for invertibility. The process defined by [4.3.3] and [4.3.4] is referred to as an ***autoregressive integrated moving average (ARIMA) process***. The reason for the term "integrated" can be found by rewriting [4.3.3] for $d=1$ as

$$z_t^{(\lambda)} = (1-B)^{-1}w_t = (1+B+B^2+\cdots)w_t = \sum_{j=0}^{\infty} w_{t-j} \qquad [4.3.5]$$

It can be seen that the $z_t^{(\lambda)}$ series can be obtained by summing or "integrating" the stationary w_t process. When the order of differencing is d then $z_t^{(\lambda)}$ is calculated by "integrating" the w_t process d times. To obtain the original z_t series from the $z_t^{(\lambda)}$ sequence, the inverse of the Box-Cox transformation in [3.4.30] is taken.

The ARIMA (p,d,q) notation is used to indicate the orders of the AR, differencing and MA operators, respectively, which are contained in the ARIMA process given by [4.3.3] and [4.3.4]. When there is no differencing (i.e., $d=0$), the set of ARIMA(p,0,q) processes is the same as the family of stationary ARMA(p,q) processes defined in Section 3.4. However, when dealing with stationary processes it has become common practice to use the term ARMA(p,q), whereas ARIMA(p,d,q) is employed whenever there is a differencing operator (i.e., $d>0$).

To demonstrate the effects of the differencing operator consider the set of ARIMA(0,d,0) models given by

$$(1-B)^d(z_t-100) = a_t \qquad [4.3.6]$$

where 100 is the mean level of the series for $d=0$ and this level disappears due to differencing when $d>0$. When $d=0$, the model is white noise. In Chapter 9, general procedures are described for simulating with white noise, ARMA, and ARIMA models. Figure 4.3.2 is a plot of 100 simulated terms from the model where the a_t's are randomly generated on a computer as being NID(0,1). It can be seen that the entries in the series appear to be uncorrelated and fluctuate about an overall mean level of 100. The same 100 a_t terms that are used for generating the sequence in Figure 4.3.2 are also employed to simulate series of length 100 for $d=1,2$ and 3. In Figure 4.3.3, a simulated sequence is shown for an ARIMA(0,1,0) model where a starting value of $z_1=100$ is utilized. Notice how the series does not fluctuate about any overall mean level and generally tends to increase in value over time. Using initial values of $z_1=100$ and $z_2=102$, a synthetic series for an ARIMA(0,2,0) model is generated in Figure 4.3.4. In that figure, the local fluctuations have largely disappeared and the sequence increases dramatically in value with increasing time. Figure 4.3.5 is a simulated trace from an ARIMA(0,3,0) model where starting values of $z_1=100$, $z_2=102$, and $z_3=104$ are employed. The simulated data increases exponentially over time and the right hand portion of the graph seems to mimic a missile trajectory.

4.3.2 Autocorrelation Function

As explained in [3.4.13] in Section 3.4.2, the theoretical ACF for an ARMA(p,q) process satisfies the difference equation

$$\phi(B)\rho_k = 0, \quad k > q \qquad [4.3.7]$$

where ρ_k is the theoretical ACF at lag k, and $\phi(B)$ is the AR operator of order p. Assuming distinct roots, the general solution for this difference equation is

$$\rho_k = A_1G_1^k + A_2G_2^k + \cdots + A_pG_p^k \qquad [4.3.8]$$

where $G_1^{-1}, G_2^{-1}, \ldots, G_p^{-1}$, are the roots of the characteristic equation $\phi(B) = 0$ and the A_i's are constants. Due to stationarity conditions, $|G_1^{-1}| > 1$ for a real root G_1^{-1}. Therefore, for increasing lag k, the term $A_iG_i^k$ damps out because $|G_i| < 1$. When all of the roots lie outside the unit circle, the theoretical ACF in [4.3.8] attenuates quickly for moderate and large lags. However, suppose that homogeneous nonstationarity is approached and at least one of the roots G_i^{-1} approaches the unit circle. This, in turn, will cause $|G_i|$ to go towards unity, $A_iG_i^k$ will not die out quickly for larger lags and, hence, ρ_k in [4.3.8] will not damp out fast for moderate and large lags.

The behaviour of the theoretical ACF for a process which is approaching homogeneous nonstationarity has some important practical implications. When the sample ACF in [2.5.9] for a given data set does not die out quickly for larger lags, this may indicate that the data should be differenced to remove homogeneous nonstationarity. For example, the sample ACF along with the 95% confidence limits is displayed in Figure 4.3.6 for the 100 simulated data points in Figure 4.3.3 which were generated by an ARIMA(0,1,0) model. Because the sample ACF attenuates slowly, this indicates the need for differencing. When the simulated sequence from Figure 4.3.3 is differenced to remove nonstationarity, the resulting sample ACF and 95% confidence limits for the differenced data are as shown in Figure 4.3.7. As expected, after differencing only white noise residuals remain. This confirms that the data were originally generated by an ARIMA(0,1,0) model.

In Figure 4.3.6, the sample ACF possesses large values at lower lags that slowly attenuate for increasing lag. However, as noted by Wichern (1973) and Roy (1977), it is not necessary that the sample ACF at the first few lags be rather large if nonstationarity is present. In certain situations, the sample ACF at low lags may in fact be relatively quite small. However, no matter how large the sample ACF values are at the first few lags, when a given data set possesses homogeneous nonstationarity the sample ACF must slowly attenuate for increasing lags.

When it is suspected that a given data set is nonstationary, the time series should be differenced just enough times to cause the sample ACF to attenuate fast for the differenced series. Following this an ARMA(p,q) model can be fitted to the differenced series which is assumed to be stationary. In practice, usually $d = 0$, 1, or 2 for ARIMA models that are fitted to many types of measured series that arise in the natural and social sciences. Furthermore, if the original data set is transformed by a Box-Cox transformation this does not eliminate the need for differencing.

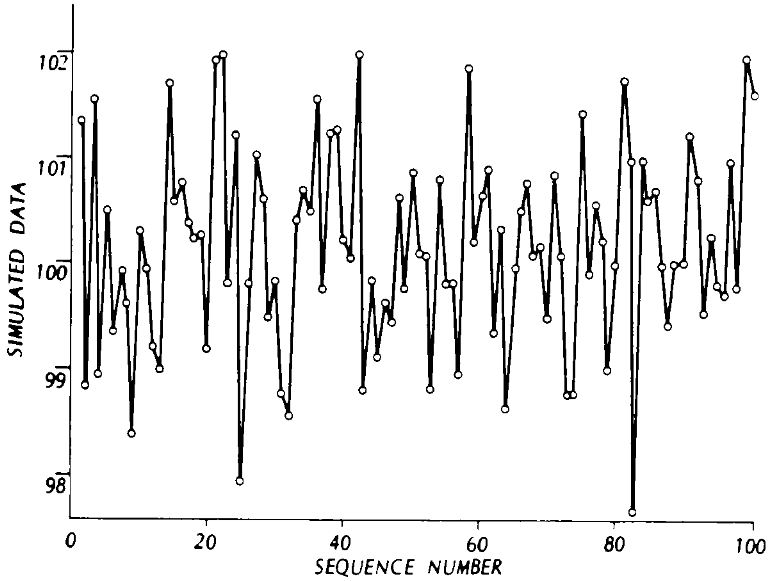


Figure 4.3.2. Simulated sequence for a white noise model.

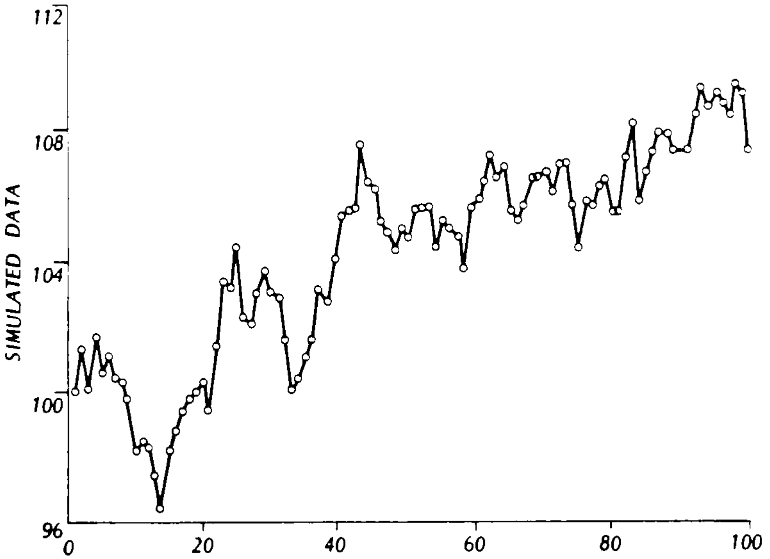


Figure 4.3.3. Simulated sequence for an ARIMA(0,1,0) model.

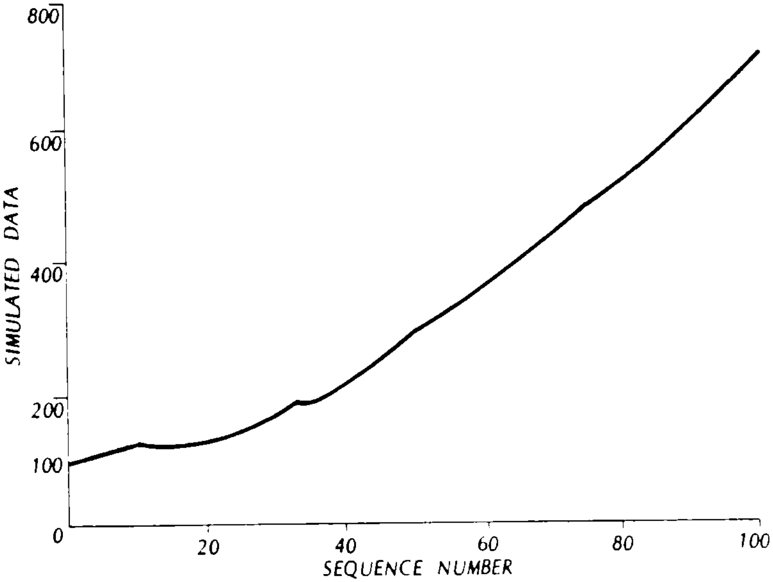


Figure 4.3.4. Simulated sequence for an ARIMA(0,2,0) model.

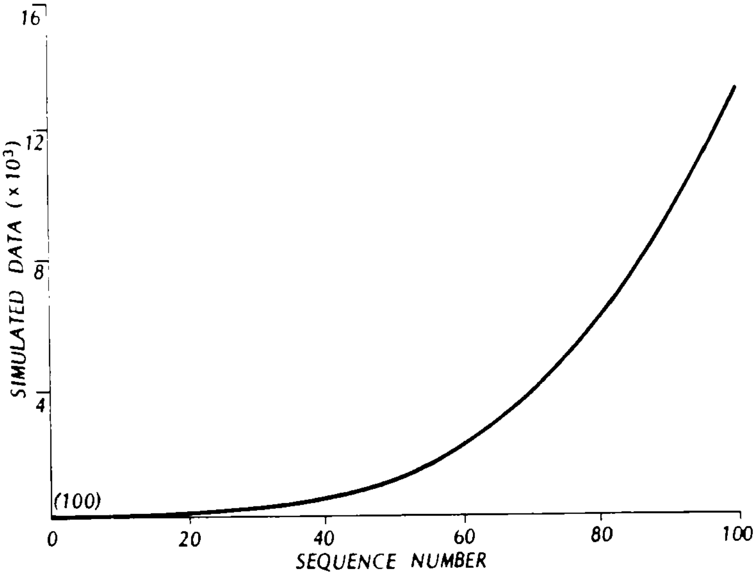


Figure 4.3.5. Simulated sequence for an ARIMA(0,3,0) model.

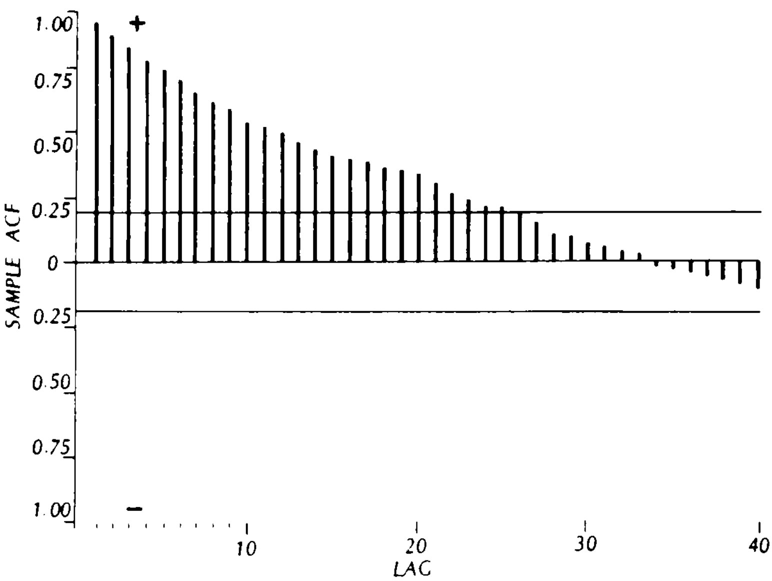


Figure 4.3.6. Sample ACF and 95% confidence limits for simulated data from an ARIMA(0,1,0) model.

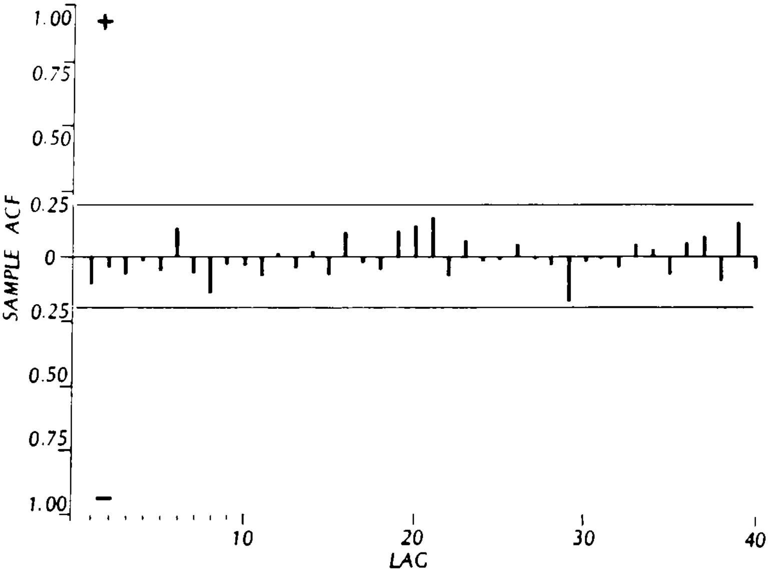


Figure 4.3.7. Sample ACF and 95% confidence limits for the differenced data generated from an ARIMA(0,1,0) model.

Rather, the transformed time series should be differenced as many times as are required to cause the sample ACF of the differenced transformed series to damp out quickly for moderate and large lags.

In certain situations, it may be difficult to ascertain whether or not a given series is nonstationary. This is because there is often no sharp distinction between stationarity and nonstationarity when the nonstationary boundary is nearby. As one or more of the roots of the characteristic equation approaches the unit circle, an ARMA process gradually changes to a nonstationary process and at the same time the corresponding theoretical ACF attenuates less quickly for increasing lags. Consequently, when examining the sample ACF for a specified data set, it is not always obvious whether or not differencing is required. If the fitted model is to be used for simulation, it may be advantageous to choose a model that does not require differencing so that the simulated data will fluctuate around an overall mean level. On the other hand, a model with a differencing operator may perform better than a stationary model when the model is used for forecasting. If employed judiciously, the Akaike information criterion (AIC) (Akaike, 1974) may be used as a guide to determine if differencing is required (see Sections 1.3.3 and 6.3).

4.3.3 Examples of Nonstationary Time Series

Annual Water Use for New York City

The annual water use for New York City is available from 1898 to 1968 in litres per capita per day (Salas and Yevjevich, 1972) and a graph of the series is portrayed in Figure 4.3.8. Because water use has tended to increase over time, the series is obviously nonstationary. The general patterns in Figure 4.3.8 are quite similar to those in Figure 4.3.3 for data that were simulated from an ARIMA(0,1,0) model. The inherent nonstationarity is also confirmed by the graph in Figure 4.3.9 of the sample ACF and 95% confidence limits of the New York water use data. The estimated ACF in Figure 4.3.9 dies off rather slowly and closely mimics the sample ACF in Figure 4.3.6 for the data that were generated from an ARIMA(0,1,0) model. When the water use data are differenced, the resulting series is white noise since all of the values of the sample ACF for the differenced data fall within the 95% confidence limits. Consequently, the annual New York water use series can be modelled by an ARIMA(0,1,0) model.

Electricity Consumption

The total annual electricity consumption for the U.S. is available from 1920 to 1970 in millions of kiloWatt - hours (United States Bureau of the Census, 1976) and a plot of the series is given in Figure 4.3.10. Due to the increase in electricity demand over time, the series is nonstationary. The behaviour of the electricity consumption series in Figure 4.3.10 closely resembles that in Figure 4.3.4 for data that were simulated from an ARIMA(0,2,0) model. As shown in Figures 4.3.11 and 4.3.12, the sample ACF's attenuate slowly for the given electricity consumption series and also the differenced series, respectively. When the series is differenced twice the nonstationarity is removed as demonstrated by the sample ACF in Figure 4.3.13. The large value at lag one indicates the need for a MA parameter in the model. At lag 9, the sample ACF just crosses the 95% confidence limits and this behaviour may be due to chance alone or could indicate the need for another parameter in the model. The sample PACF in Figure 4.3.14 for the electricity consumption data may be interpreted as attenuating quickly at the first few lags due to the need for a MA component. Based upon this identification information, the most appropriate model to the electricity consumption data is an ARIMA(0,2,1) model. Moreover, when one

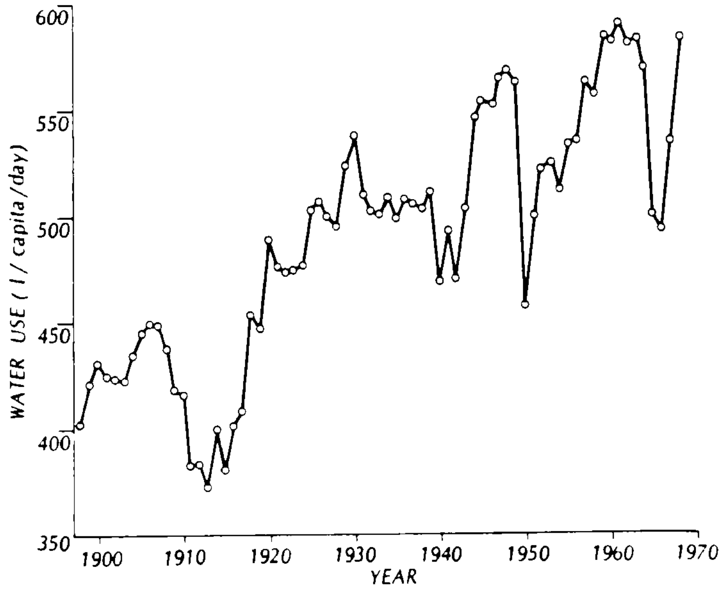


Figure 4.3.8. Annual water use for New York City.

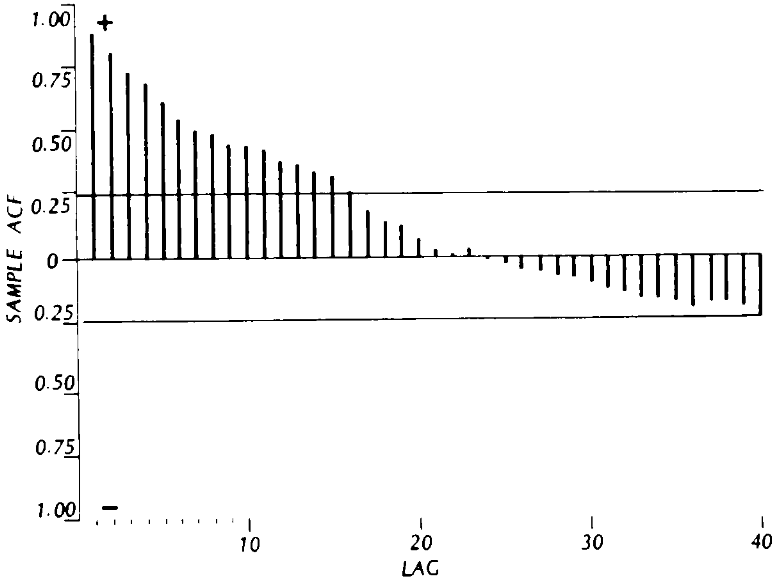


Figure 4.3.9. Sample ACF and 95% confidence limits for the annual water use of New York City.

obtains a maximum likelihood estimate (see Section 6.2) of the Box-Cox parameter λ in [3.4.30], the estimated value is $\lambda = 0.533$, which is essentially a square root transformation (i.e. $\lambda \approx 0.5$). The need for a data transformation can be visually detected by examining the graphs of a smoothing procedure which divides the original graph of the electricity demand series into smooth and rough plots (see Section 22.3).

Beveridge Wheat Price Index

The annual Beveridge wheat price index series which is available from 1500 to 1869 (Beveridge, 1921) is shown in Figure 4.3.15. This series could be closely related to climatic conditions and, therefore, may be of interest to hydrologists and climatologists. For example, during years when the weather is not suitable for abundant grain production, the price of wheat may greatly escalate. If a model can be developed that relates a given hydrologic time series to the Beveridge wheat price indices, this model could be employed to extend the hydrologic record if it were shorter than the other data set (see Sections 17.5.4, 18.5.2 and 19.3.2).

From a plot of the Beveridge wheat price indices in Figure 4.3.15 for the period from 1500 to 1869, it can be seen that the series is nonstationary. Both the level and variance of the time series are increasing over time. A change in variance over time of the original data would eventually be mirrored by variance that is not constant in the residuals of the model fitted to the data. To rectify the situation from the start, natural logarithms are taken of the series so that the variance changes are not as drastic as those shown in Figure 4.3.15. The sample ACF is given for the logarithmic series from 1500 to 1869 in Figure 4.3.16. Because the sample ACF attenuates very slowly for increasing lag, the logarithmic data set should be differenced to remove the inherent nonstationarity. Figure 4.3.17 is a plot of the sample ACF for the differenced logarithmic data along with the 95% confidence limits where it is assumed that the estimated ACF is not significantly different from zero after lag 3. In addition to the large values at low lags, the sample ACF just touches the 95% confidence limits at lag 8. The graph of the sample PACF and 95% confidence limits for the differenced logarithmic series is presented in Figure 4.3.18. A rather large value of the estimated PACF exists at lag 2 while there is a value that crosses the 95% confidence limits at lag 8. Therefore, an AR operator that includes parameters at low lags and also lag 8, may be required in a model that is fitted to the data. After considering a number of possible models, it is found that the most appropriate model to fit to the logarithmic series is a constrained ARIMA(8,1,1) model where ϕ_3 to ϕ_7 are not included in the AR operator.

4.3.4 Three Formulations of the ARIMA Process

In Section 3.4.3, it is shown how the difference equation for the ARMA(p,q) process in [3.4.4] can also be written in the random shock form as an infinite MA process in [3.4.18] or else in the inverted form as an infinite AR process in [3.4.25]. The results in Section 3.4.3 also hold for the stationary w_t process in [4.3.4] which is made stationary by differencing the nonstationary $z_t^{(\lambda)}$ process in [4.3.3]. By using similar procedures, the ARIMA difference equation for the nonstationary $z_t^{(\lambda)}$ process can also be conveniently expressed in either the random shock or inverted forms.

Treating $\phi(B)$, $\theta(B)$, and $(1-B)^d$ as algebraic operators, the *random shock form* of the ARIMA process is

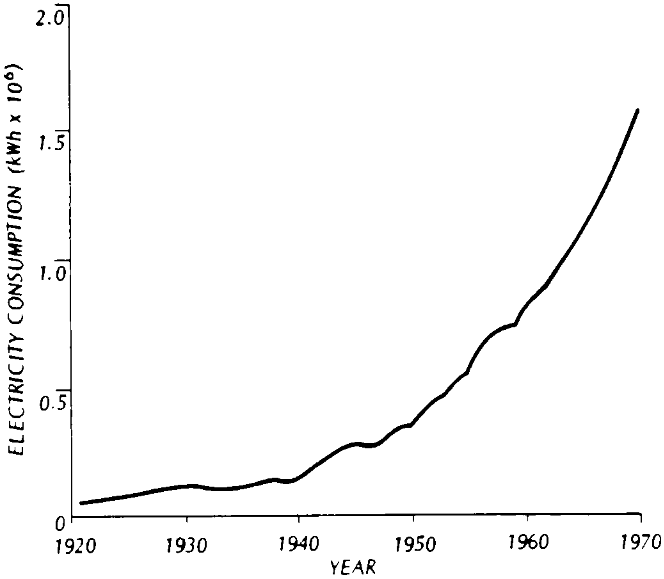


Figure 4.3.10. Total annual electricity consumption in the U.S.A.

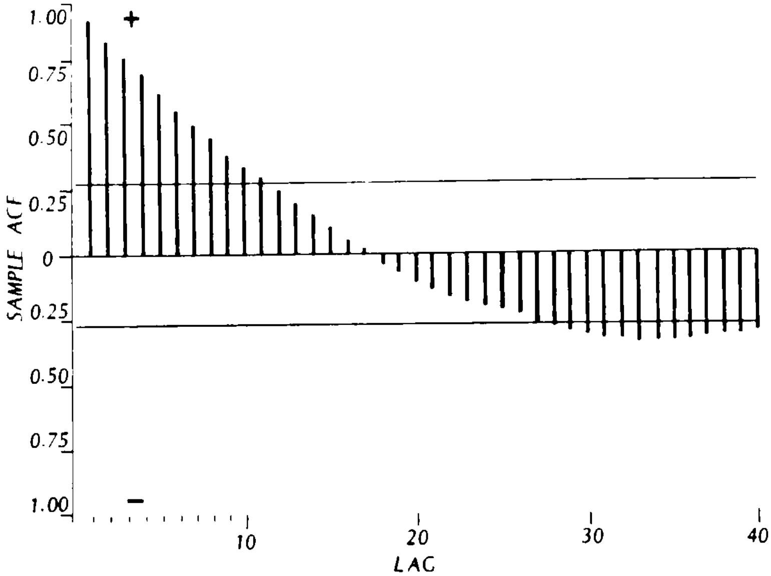


Figure 4.3.11. Sample ACF and 95% confidence limits for the annual American electricity consumption.

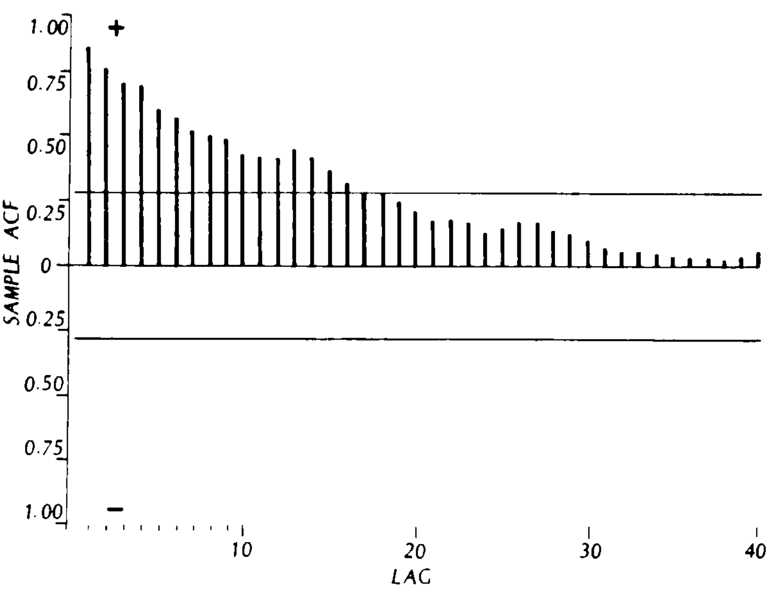


Figure 4.3.12. Sample ACF and 95% confidence limits for the differenced annual American electricity consumption series.

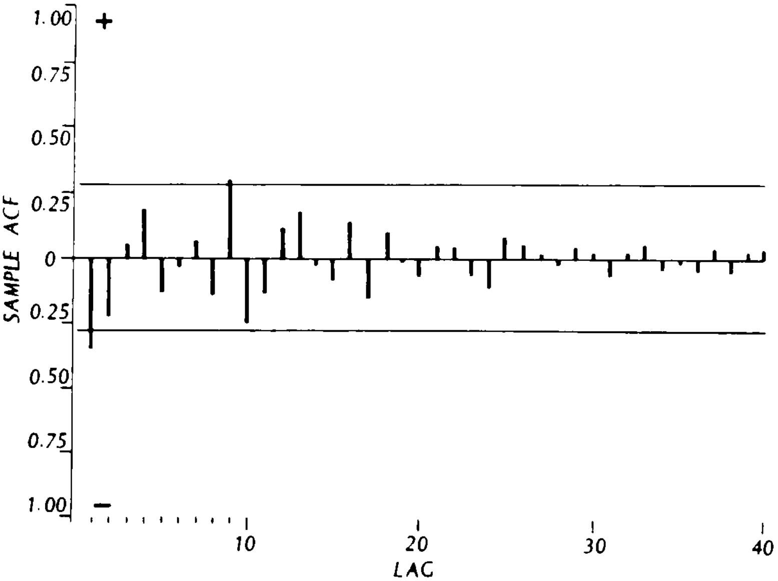


Figure 4.3.13. Sample ACF and 95% confidence limits for the annual American electricity consumption series that is differenced twice.

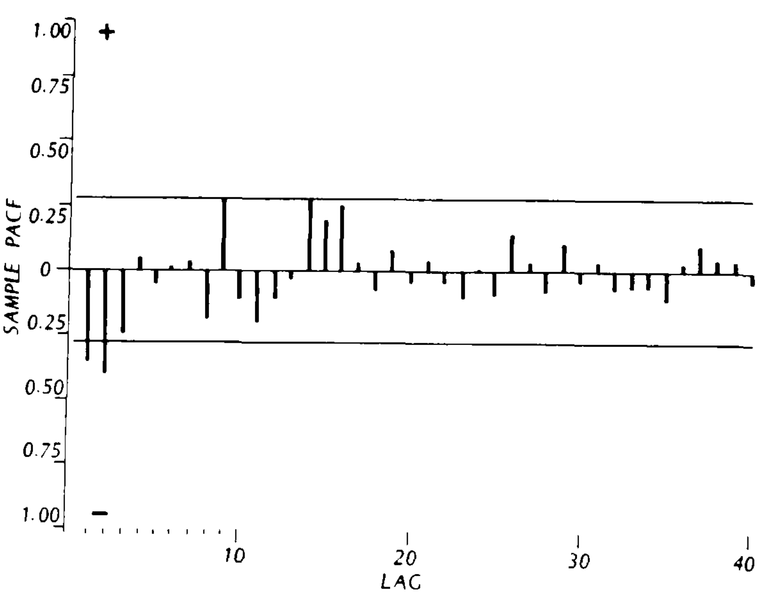


Figure 4.3.14. Sample PACF and 95% confidence limits for the annual American electricity consumption series that is differenced twice.

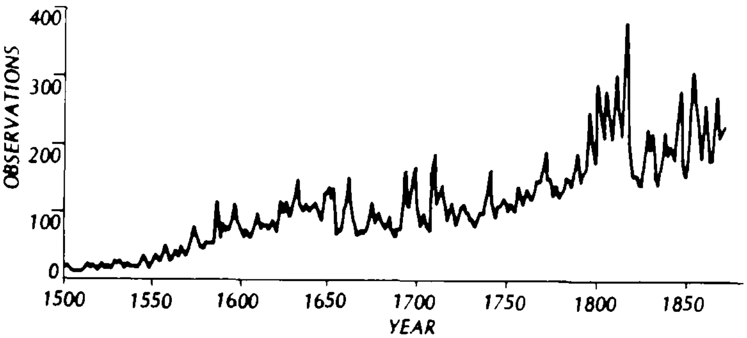


Figure 4.3.15. Beveridge wheat price indices from 1500 to 1869.

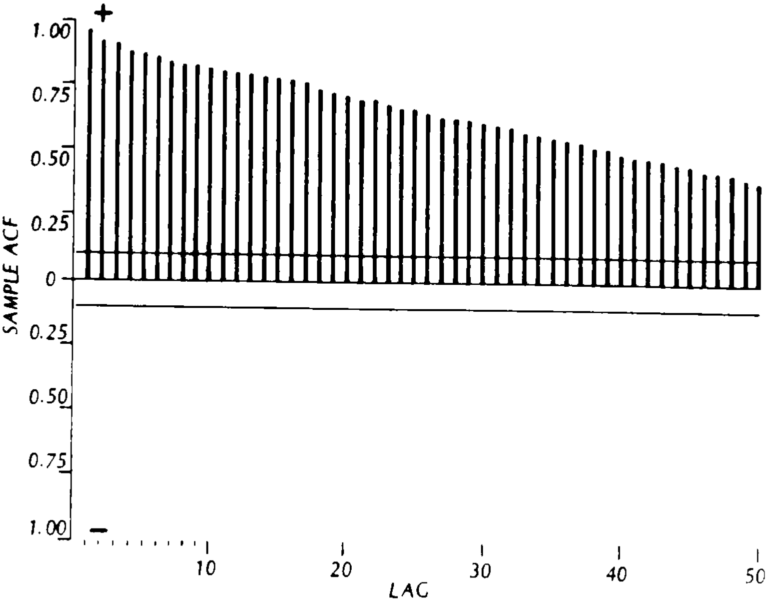


Figure 4.3.16. Sample ACF and 95% confidence limits for the logarithmic Beveridge wheat price index series.

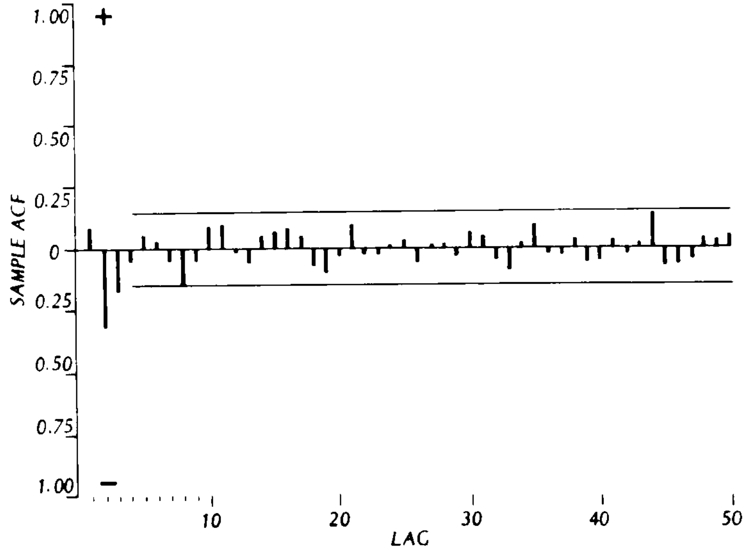


Figure 4.3.17. Sample ACF and 95% confidence limits for the differenced logarithmic Beveridge wheat price index series when ρ_k is zero after lag 3.

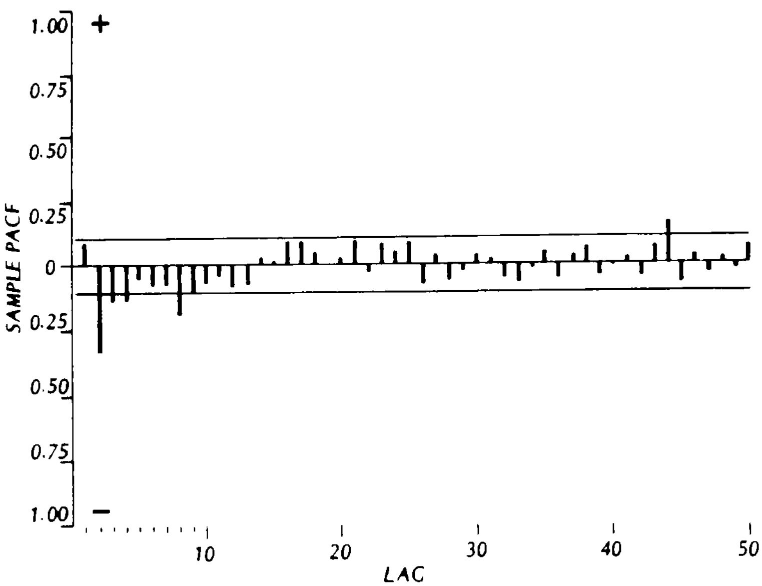


Figure 4.3.18. Sample PACF and 95% confidence limits for the differenced logarithmic Beveridge wheat price index series.

$$
\begin{aligned}
z_t^{(\lambda)} &= [\phi(B)(1-B)^d]^{-1}\theta(B)a_t \\
&= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\
&= (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_t \\
&= \psi(B)a_t \qquad [4.3.9]
\end{aligned}
$$

where $\psi(B)$ is the *random shock or infinite MA operator* and ψ_i is the ith parameter or weight of $\psi(B)$. To develop a relationship for ascertaining the ψ parameters, first multiply [4.3.9] by $\phi(B)(1-B)^d$ to obtain

$$\phi(B)(1-B)^d z_t^{(\lambda)} = \phi(B)(1-B)^d \psi(B)a_t$$

From [4.3.3] and [4.3.4], $\theta(B)a_t$ can be exchanged for $\phi(B)(1-B)^d z_t^{(\lambda)}$ in the previous equation to get

$$\theta(B) = \phi(B)(1-B)^d \psi(B) \qquad [4.3.10]$$

The ψ weights can be readily determined by expressing [4.3.10] as

$$\phi(B)(1-B)^d \psi_k = -\theta_k \qquad [4.3.11]$$

where B operates on k, $\psi_0 = 1$, $\psi_k = 0$ for $k < 0$ and $\theta_k = 0$ if $k > q$. As is done for the examples in Section 3.4.3, the ψ weights can be recursively calculated by solving [4.3.11] for $k = 1,2,\ldots,q'$, where q' is the number of ψ parameters that are required.

In order to write the *inverted form* of the process, the ARIMA process is reformulated as

$$\begin{aligned} a_t &= \theta(B)^{-1}\phi(B)(1-B)^d z_t^{(\lambda)} \\ &= z_t^{(\lambda)} - \pi_1 z_{t-1}^{(\lambda)} - \pi_2 z_{t-2}^{(\lambda)} - \cdots \\ &= (1 - \pi_1 B - \pi_2 B^2 - \cdots)z_t^{(\lambda)} \\ &= \pi(B)z_t^{(\lambda)} \end{aligned} \qquad [4.3.12]$$

where $\pi(B)$ is the *inverted or infinite AR operator* and π_i is the ith parameter or weight of $\pi(B)$. To determine a relationship for calculating the π parameters, multiply [4.3.12] by $\theta(B)$ to get

$$\theta(B)a_t = \theta(B)\pi(B)z_t^{(\lambda)}$$

By employing [4.3.3] and [4.3.4], $\phi(B)(1-B)^d z_t^{(\lambda)}$ can be substituted for $\theta(B)a_t$ in the above equation to obtain

$$\phi(B)(1-B)^d = \theta(B)\pi(B) \qquad [4.3.13]$$

The π coefficients can be easily ascertained by expressing the above equation as

$$\theta(B)\pi_k = (1-B)^d \phi_k \qquad [4.3.14]$$

where $\pi_0 = -1$ and $\phi_0 = -1$ when using [4.3.14] to calculate π_k for $k > 0$, $\pi_k = 0$ for $k < 0$, and $\phi_k = 0$ if $k > p$ or $k < 0$. By solving [4.3.14] for $k = 1,2,\ldots,p'$, where p' is the number of π parameters that are needed, the π weights can be recursively calculated in the same fashion as the examples in Section 3.4.3.

An interesting property of the π weights is when $d \geq 1$ the parameters in the inverted operator sum to unity. This fact can be proven by substituting $B = 1$ into [4.3.13]. In that equation, $\phi(1)$ and $\theta(1)$ are not zero since the roots of the characteristic equations for the AR and MA operators lie outside the unit circle. However, $(1-B)^d = 0$ for $B = 1$ and therefore [4.3.13] reduces to

$$\pi(1) = 0$$

or

$$\sum_{j=1}^{\infty} \pi_j = 1 \qquad [4.3.15]$$

Consequently, for $d \geq 1$ equation [4.3.12] can be written as

$$z_t^{(\lambda)} = \sum_{j=1}^{\infty} \pi_j z_{t-j}^{(\lambda)} + a_t \qquad [4.3.16]$$

where the summation term on the right hand side constitutes a weighted average of the previous values of the process.

4.4 INTEGRATED MOVING AVERAGE PROCESSES

In Section 4.3.3, it was found that the most appropriate model to fit to the total annual electricity consumption in the U.S.A. is an ARIMA(0,2,1) model. When modelling time series from economics and other fields of study, it often turns out that ARIMA models are needed where $p = 0$ and both d and q are greater than zero. Because no AR operator is present, an ARIMA(0,d,q) process is often referred to as an *integrated moving average (IMA) process* and is denoted by IMA(0,d,q). For a detailed description of IMA processes, the reader may wish to refer to the book of Box and Jenkins (1976, Ch. 4, pp. 103-114).

A special case of the IMA(0,d,q) family of processes is the IMA(0,1,1) process given by

$$(1 - B)z_t^{(\lambda)} = (1 - \theta_1 B)a_t$$

or

$$z_t^{(\lambda)} = z_{t-1}^{(\lambda)} + a_t - \theta_1 a_{t-1}$$

Keeping in mind that the data, z_t, may require a Box-Cox transformation, the above equation can be more conveniently written by dropping the λ superscript and writing it as

$$z_t = z_{t-1} + a_t - \theta_1 a_{t-1} \qquad [4.4.1]$$

The minimum mean square error forecasts (see Section 8.2) obtained from an IMA(0,1,1) process, are the same forecasts that are produced when using *single exponential smoothing* [see, for example, Gilchrist (1976, p. 108)]. Because exponential smoothing has been used extensively for forecasting economic time series [see, for instance, Makridakis and Wheelwright (1978) and Gilchrist (1976)], the IMA(0,1,1) process has received widespread attention. Important original research regarding the optimal properties of exponentially weighted forecasts is given by Muth (1960).

To appreciate the inherent structure of the IMA(0,1,1) process in [4.4.1], the random shock form of the process in [4.3.9] is useful. The ψ coefficients can be obtained by employing [4.3.11] for positive values of k. For $k = 1$

$$(1 - B)\psi_1 = -\theta_1 \text{ or } \psi_1 - \psi_0 = -\theta_1$$

But $\psi_0 = 1$ and, therefore, $\psi_1 = 1 - \theta_1$.

When $k = 2$

$$(1 - B)\psi_2 = 0 \text{ or } \psi_2 - \psi_1 = 0$$

Therefore, $\psi_2 = \psi_1 = (1 - \theta_1)$.

For $k = 3$

$$(1 - B)\psi_3 = 0 \text{ or } \psi_3 - \psi_2 = 0$$

Therefore, $\psi_3 = \psi_2 = (1 - \theta_1)$.

In general,

$$(1-B)\psi_k = 0 \text{ or } \psi_k - \psi_{k-1} = 0$$

Therefore, $\psi_k = \psi_{k-1} = \cdots = \psi_1 = (1-\theta_1)$.

By substituting for the ψ parameters into [4.3.9], the random shock form of the model is

$$z_t = (1-\theta_1)\sum_{j=1}^{\infty} a_{t-j} + a_t \qquad [4.4.2]$$

From [4.4.2] it can be seen that the present value of the process depends upon the current random shock, a_t, plus the summation of an equal weighting of all previous disturbances. Consequently, part of the random shock in any period has a permanent effect due to the weight, $(1-\theta_1)$, while the rest affects the system only in the current time period.

The inverted form of the process can be employed for understanding the properties of an IMA(0,1,1) process. By examining [4.3.14] for positive values of k, the π coefficient can be ascertained. For $k = 1$

$$(1-\theta_1 B)\pi_1 = (1-B)\phi_1 \text{ or } \pi_1 - \theta_1\pi_0 = \phi_1 - \phi_0$$

But $\pi_0 = \phi_0 = -1$ when determining the π weights and $\phi_1 = 0$ since $p = 0$. Therefore, $\pi_1 = 1 - \theta_1$. When $k = 2$

$$(1-\theta_1 B)\pi_2 = (1-B)\phi_2 \text{ or } \pi_2 - \theta_1\pi_1 = \phi_2 - \phi_1$$

Because no AR parameters are present in the IMA(0,1,1) process, $\phi_1 = \phi_2 = 0$ and, therefore, $\pi_2 = \theta_1\pi_1 = \theta_1(1-\theta_1)$. For $k = 3$

$$(1-\theta_1 B)\pi_3 = (1-B)\phi_3 \text{ or } \pi_3 - \theta_1\pi_2 = \phi_3 - \phi_2 = 0$$

Hence, $\pi_3 = \theta_1\pi_2 = \theta_1^2(1-\theta_1)$.

In general, the π coefficient at lag k is determined by

$$(1-\theta_1 B)\pi_k = (1-B)\phi_k \text{ or } \pi_k - \theta_1\pi_{k-1} = \theta_1^{k-1}(1-\theta_1)$$

By substituting for the π parameters into [4.3.16], the inverted form of the process is

$$z_t = (1-\theta_1)\sum_{j=1}^{\infty}\theta_1^{j-1} z_{t-j} + a_t \qquad [4.4.3]$$

The summation term on the right hand side of [4.4.3] constitutes an *exponentially weighted moving average (EWMA)* of the previous values of the process and is denoted as

$$\bar{z}_{t-1}(\theta_1) = (1-\theta_1)\sum_{j=1}^{\infty}\theta_1^{j-1} z_{t-j} \qquad [4.4.4]$$

The weights in [4.4.4] are formed by the sequence of π parameters given by $(1-\theta_1),(1-\theta_1)\theta_1,(1-\theta_1)\theta_1^2,(1-\theta_1)\theta_1^3,\ldots$. When θ_1 has a value of zero, the IMA(0,1,1) process in [4.4.2] reduces to an IMA(0,1,0) process where $\pi_1 = 1$ and $\pi_k = 0$ for $k>1$. As the value of θ_1 approaches unity, the π weights attenuate more slowly and the EWMA in [4.4.4] stretches further into the past of the process. When θ_1 is equal to one, the MA and differencing operators cancel in [4.4.1] and the process is a white noise IMA(0,0,0) process.

From its definition in [4.4.4], the recursion formula for the EWMA can be written as

$$\bar{z}_t(\theta_1) = (1 - \theta_1)z_t + \theta_1\bar{z}_{t-1}(\theta_1) \qquad [4.4.5]$$

This expression is what is employed for obtaining forecasts using single exponential smoothing [see, for example, Makridakis and Wheelwright (1978, Ch. 5)]. Although the IMA(0,1,1) process possesses no mean due to the fact that it is nonstationary, the EWMA in [4.4.4] can be regarded as being the location or level of the process. From [4.4.5] it can be seen that each new level is calculated by interpolating between the new observation and the previous level. When θ_1 is equal to zero, the process is actually an IMA(0,1,0) process and the current level in [4.4.5] would be solely due to the present observation. If θ_1 were close to unity, the current level, $\bar{z}_t(\theta_1)$, in [4.4.5] would depend heavily upon the previous level, $\bar{z}_{t-1}(\theta_1)$, while the current observation, z_t, would be given a small weight of $(1 - \theta_1)$.

Muth (1960) suggests an intuitive approach for interpreting the generation of the single exponential smoothing procedure or equivalently the IMA(0,1,1) process. From [4.4.3] and [4.4.4]

$$z_t = \bar{z}_{t-1}(\theta_1) + a_t$$

By substituting [4.4.3] into [4.4.5] it turns out that

$$\bar{z}_t(\theta_1) = \bar{z}_{t-1}(\theta_1) + (1 - \theta_1)a_t \qquad [4.4.6]$$

The first of the previous two equations demonstrates how the current value z_t is produced by the level of the system at time $t-1$ plus a random shock added at time t. However, [4.4.6] shows that only a proportion, $(1 - \theta_1)$, of the innovation has a lasting influence by being absorbed into the current level of the process.

4.5 DIFFERENCING ANALOGIES

When dealing with discrete data, the differencing operator $\nabla^d = (1 - B)^d$ can be employed to remove homogeneous nonstationarity. It turns out that the differencing operator is analogous to differentiation when continuous functions are being studied. Consider, for example, a discrete process which is defined by

$$z_t = \begin{cases} a_t & \text{for } t < T \\ c + a_t & \text{for } t \geq T \end{cases} \qquad [4.5.1]$$

where c is a constant which reflects a local level for $t \geq T$. When a_t is assumed to be IID$(0,\sigma_a^2)$, the mean level of the z_t process before time T is zero while the mean of the process is c for $t \geq T$. The effect of differencing the data once is to remove the local level due to the constant c in [4.5.1]. For $t>T$ the differenced series is calculated as

$$\nabla z_t = (1-B)z_t = z_t - z_{t-1} = (c + a_t) - (c + a_{t-1})$$
$$= a_t - a_{t-1}$$

The above operation is analogous to taking the first derivative of a continuous function of time which is given as

$$y = \begin{cases} 0 & \text{for } t < T \\ c & \text{for } t \geq T \end{cases}$$

The derivative $\frac{dy}{dt}$ is of course zero for $t > T$ and the local level drops out due to differentiation.

Next, consider the analogous effects of differencing operators of order two for the discrete case and second order derivatives for a continuous function. Suppose that a discrete process is given as

$$z_t = c + bt + a_t \qquad [4.5.2]$$

where b and c are constants. The term, $(c + bt)$, forms a linear deterministic trend while the white noise, a_t, constitutes the probabilistic component of the process, z_t. By using a differencing operator of order one, the constant c in [4.5.2] can be removed as is shown by

$$\nabla z_t = (1-B)z_t = z_t - z_{t-1}$$
$$= (c + bt + a_t) - (c + b(t-1) + a_{t-1})$$
$$= b + a_t - a_{t-1}$$

By employing a differencing operator of order two, the entire deterministic trend can be eliminated.

$$\nabla^2 z_t = \nabla(\nabla z_t)$$
$$= (b + a_t - a_{t-1}) - (b + a_{t-1} - a_{t-2})$$
$$= a_t - 2a_{t-1} + a_{t-2}$$

For the continuous case, a function of t may be given as

$$y = c + bt$$

The value of the first derivative is $\frac{dy}{dt} = b$ while $\frac{d^2y}{dt^2} = 0$. Hence, the first order derivative removes the intercept, c, while the second order derivative completely eliminates the linear function.

4.6 DETERMINISTIC AND STOCHASTIC TRENDS

The component $c + bt$ in [4.5.2] is an example of a deterministic linear trend component. In general, the *deterministic trend* component could be any function $f(t)$ and after the trend component is removed from the time series being studied, the residual could be modelled by an appropriate stochastic model. For example, suppose that the series is transformed by a Box-Cox transformation and following this a trend component $f(t)$ and perhaps also an overall mean level μ are subtracted from the transformed series. If the resulting series were modelled by an ARMA(p,q) model, the model would be written as

$$\phi(B)(z_t^{(\lambda)} - f(t) - \mu) = \theta(B)a_t \qquad [4.6.1]$$

This type of procedure is similar to what is used with the deseasonalized models in Chapter 13. Due to the annual rotation of the earth around the sun, there is a physical justification for including a sinusoidal deterministic component when modelling certain kinds of natural seasonal time series. Consequently, the data are deseasonalized by removing a deterministic sinusoidal component and following this the resulting nonseasonal series is modelled using an ARMA(p,q) model.

The model in [4.6.1] possesses a deterministic trend component. In certain types of series with linear trends, the trends may not be restricted to occur at a specified time nor have approximately the same slope or duration. Rather, the trends may occur stochastically and there may be no physical basis for justifying the use of a deterministic trend. As was demonstrated in the previous section, a differencing operator of order two could account for linear trends if they were known or expected to be present. Consequently, to allow for stochastic linear trends, the series which may have first been changed by a Box-Cox transformation could be differenced twice before an ARMA(p,q) model is fitted. In general, *stochastic trends* of order $d - 1$ are automatically incorporated into the ARIMA(p,d,q) model

$$\phi(B)\nabla^d z_t^{(\lambda)} = \theta(B)a_t \qquad [4.6.2]$$

In certain instances, it may not be clear as to whether or not one should include a deterministic trend component in the model. Recall that the w_t sequence in [4.3.3] is assumed to have a mean of $\mu_w = 0$ after the $z_t^{(\lambda)}$ series is differenced d times. However, if the estimated mean of μ_w were significantly different from zero this may indicate that differencing cannot remove all of the nonstationarity in the data and perhaps a deterministic trend is present. When estimating the parameters of an ARIMA model which is fit to a given data set, the MLE (maximum likelihood estimate) $\hat{w}$ of μ_w can be obtained (see Chapter 6). Because a MLE possesses a limiting normal distribution, by using the estimated SE (standard error) and subjectively choosing a level of significance, significance testing can be done for the estimated model parameter. For instance, if the absolute value of $\hat{w}$ is less than twice its SE, it can be argued that $\hat{w}$ is not significantly different from zero and should be omitted from the model. Likewise, when estimating the sample ACF, the mean of the differenced series can be set equal to zero when it is thought that a deterministic trend component is not present. For the sample ACF's of differenced series that are examined in this chapter (see, for instance, Figure 4.3.13), it is assumed that the mean of the differenced series is zero. On the other hand, when a deterministic trend is contained in the data, the mean of the differenced series should be removed when estimating the sample ACF. This will preclude the masking of information in the plot of the sample ACF that can assist in identifying the AR and MA parameters that are required in a model which is fitted to the series.

4.7 CONCLUSIONS

As demonstrated by the interesting applications of Section 4.3.3, the ARIMA(p,d,q) model in [4.3.4] is capable of modelling a variety of time series containing stochastic trends. The first step in modelling a given time series is to ascertain if the data are nonstationary. If, for example, the sample ACF attenuates slowly, this may indicate the presence of nonstationarity and the need for differencing to remove it. Subsequent to obtaining a stationary series, an ARMA(p,q) model can be fitted to the differenced data. If the model residuals are not homoscedastic (i.e., have constant variance) and/or normally distributed, the original time series can be transformed using the Box-Cox transformation in [3.4.30] in order to rectify the situation. Following this, an ARIMA(p,d,q) model can be fitted to the transformed time series by using the procedure just described for the untransformed one.

For the ARIMA(p,d,q) model in [4.3.4], it is assumed that *d* can only have values that are non-negative integers. A generalization of the ARIMA model is to allow *d* to be a real number. For a specified range of the parameter *d*, the resulting process will possess long memory (see Section 2.5.3 for a definition of long and short memory processes) and, consequently, this process is discussed in more detail with other long memory processes in Part V. As explained in Chapter 11 in Part V, when *d* is allowed to take on real values, the resulting model is referred to as a fractional ARMA or FARMA process. However, before presenting some long memory processes in Part V, the identification estimation, and diagnostic check stages of model construction are described in Part III for use with the stationary and nonstationary linear time series models of Part II. Many of the model building tools of part III are modified and extended for employment with the FARMA models of Chapter 11 as well as the many other types of models presented in the book and listed in Table 1.6.2.

PROBLEMS

4.1 List the names of five types of yearly time series which you expect would be nonstationary. Give reasons for your suspicions. Refer to a journal such as Water Resources Bulletin, Stochastic Hydrology and Hydraulics, Journal of Hydrology, Environmetrics or Water Resources Research and find three examples of yearly nonstationary series. How did the authors of the paper, in which a given series appeared, model the nonstationarity?

4.2 In Section 4.3.1, it is pointed out that a time series should be differenced just enough times to remove homogeneous nonstationarity. What happens if you do not difference the series enough times before fitting an ARMA model to it? What problems can arise if the series is differenced too many times?

4.3 By referring to the paper of Roy (1977), explain why the values of the sample ACF at the first few lags do not have to be large if nonstationarity is present.

4.4 An ARIMA(1,2,1) model is written as

$$(1-B)^2(1-0.8B)z_t = (1-0.5B)a_t$$

Write this model in the random shock and inverted forms. Determine at least seven random shock and inverted parameters.

4.5 For the model in question 4.4, simulate a sequence of 20 values assuming that the innovations are NID(0,1). Simulate another sequence of 20 values using innovations that are NID(0,25). Plot the two simulated sequences and compare the results. To obtain each synthetic data set, you can use a computer programming package such as the McLeod-Hipel Time Series package referred to in Section 1.7. Moreover, you may wish to examine synthetic data generated from other types of ARIMA models.

4.6 Write down the definition of a single exponential smoothing model. Show why the forecasts from this model are the same as the minimum mean square error forecasts obtained from an IMA(0,1,1) model.

4.7 Give the definition of a random walk process. What is the relationship between a random walk process and an IMA(0,1,1) process?

4.8 For each of the series found in question 4.1, explain what type of trend do you think is contained in the data? How would you model each series?

4.9 Outline the approaches that Pandit and Wu (1983) suggest for modelling stochastic and deterministic trends in Chapters 9 and 10, respectively, in their book. Compare these to the procedures described in Section 4.6 and elsewhere in this book.

4.10 Describe the procedure of Abraham and Wu (1978) for detecting the need for a deterministic component when modelling a given time series. Discuss the advantages and drawbacks of their approach.

REFERENCES

DATA SETS

Beveridge, W. H. (1921). Weather and harvest cycles. *Economics Journal*, 31:429-552.

Salas, J. D. and Yevjevich, V. (1972). Stochastic structure of water use time series. Hydrology Paper No. 52, Colorado State University, Fort Collins, Colorado.

United States Bureau of the Census (1976). *The Statistical History of the United States from Colonial Times to the Present*. Washington, D.C.

FORECASTING

Gilchrist, W. (1976). *Statistical Forecasting*. Wiley, New York.

Makridakis, S. and Wheelright, S. C. (1978). *Forecasting - Methods and Applications*. Wiley, New York.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55:299-306.

TIME SERIES ANALYSIS

Abraham, B. and Box, G. E. P. (1978). Deterministic and forecast-adaptive time-dependent models. *Applied Statistics*, 27(2):120-130.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716-723.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Pandit, S.M. and Wu, S.M. (1983). *Time Series and System Analysis with Applications*. Wiley, New York.

Roy, R. (1977). On the asymptotic behaviour of the sample autocovariance function for an integrated moving average process. *Biometrika*, 64(2):419-421.

Wichern, D. W. (1973). The behaviour of the sample autocorrelation function for an integrated moving average process. *Biometrika*, 60(2):235-239.

PART III

MODEL CONSTRUCTION

In Part II, a range of flexible types of nonseasonal models are defined and some useful theoretical properties of these models are presented. More specifically, **Chapter 3 describes AR, MA, and ARMA models,** which can be applied to stationary nonseasonal time series. **Chapter 4 deals with ARIMA models** which can be fitted to nonstationary nonseasonal time series. Within Chapters 3 and 4, it is pointed out that one can decide upon which particular kind of model to fit to a given data set by selecting a model whose theoretical properties are compatible with the statistical properties of the time series. For example, if the sample ACF of the data only has values which are significantly different from zero at lags one and two, one may wish to fit a MA(2) model to the time series because it is known that the theoretical ACF of a MA(2) model is exactly equal to zero after lag 2 (see Section 3.3.2). Although the foregoing and other aspects of how to fit models to data are described in Part II, there are many other valuable tools that are required in order to use the theoretical models of Part II in practical applications. Consequently, **the major objectives of Part III are to present a comprehensive methodology for applying theoretical models to actual time series and to describe a wide range of useful tools for implementing this methodology in practice.**

The overall methodology to fitting models to data is referred to as **model construction**. As portrayed in Figure III.1 and also Figure 1.3.1, model construction consists of identification, estimation and diagnostic checking. Before starting these three model development stages, one must decide upon which **families of models** should be considered for fitting to a time series. If, for example, one wishes to determine the most appropriate model to fit to a stationary nonseasonal time series, then the ARMA(p,q) models defined in Chapter 3 can be entertained. At the **identification stage** the most suitable models to fit to the data can be selected by examining various types of graphs. Although sometimes it is possible to choose the best model based solely upon identification results, in practice often it is not obvious which model is most appropriate, and hence two or three models must be tentatively entertained. For the case of ARMA(p,q) models, one must determine the number of AR and MA parameters which may be needed in the model. Efficient estimates of the model parameters can be obtained at the **estimation stage** by employing the method of maximum likelihood. Following this, the fitted models can be checked for possible inadequacies. If the **diagnostic tests** reveal serious model anomolies for the fitted model which appears to be most appropriate, then the necessary **model modifications** can be made by repeating the three stages of model development. As shown in Figure III.1, the model which is ultimately selected can then be used for **application purposes**.

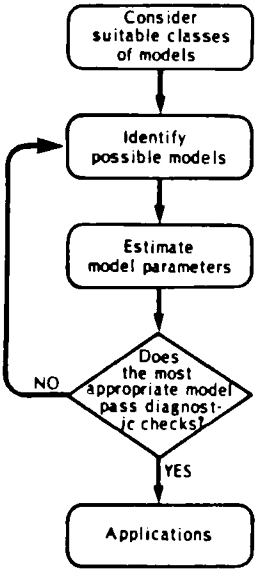

Consider suitable classes of models
Identify possible models
Estimate model parameters
Does the most appropriate model pass diagnostic checks?
NO
YES
Applications

Figure III.1. Model construction.

Specific model construction tools that can be used with the theoretical models of Part II are described in Part III. In particular, useful identification, estimation and diagnostic check techniques are presented in Chapters 5 to 7, respectively. When applying the many other kinds of models described later in this book in Parts V to IX, one can follow the same basic methodology given in Figure III.1. As a matter of fact, many of the methods and algorithms presented in Part III, or appropriate variations thereof, can be used as part of the model building for the other kinds of models in this book.

Adherence to the three phases of model development is analogous to a client obtaining a tailor-made suit from a merchant. When the customer enters the tailor's shop, he must decide upon the style and colour of the suit that he wants and the tailor then "identifies" the pattern for the suit by taking appropriate measurements of his client. At the next stage, the tailor cuts out his pattern on a bolt of cloth and sews the suit together. Finally, the customer determines if the suit fits properly by trying on the new clothing and viewing himself in front of a mirror. If alterations are required, the tailor can take suitable measurements and then make the necessary adjustments to the suit. This procedure can be repeated until the client is satisfied with his new attire. The details of the tailor's three step approach to doing business are now described in the next three chapters.

CHAPTER 5

MODEL IDENTIFICATION

5.1 INTRODUCTION

Because observations measured from natural phenomena possess an inherent probabilistic structure, time series models are employed for modelling water resources and environmental systems. The purpose of this chapter is to present informative graphical methods for identifying the most appropriate type of ARMA (Chapter 3) or ARIMA (Chapter 4) model to fit to a specified nonseasonal sequence of observations. Following a discussion of modelling philosophies in the next section, some useful graphical techniques are described in Section 5.3. Applications presented in Section 5.4, as well as identification examples introduced in Part II, clearly demonstrate that the identification methods can be conveniently applied in practice to natural time series. Prior to the conclusions, other identification methods for designing ARMA and ARIMA models are discussed in Section 5.5.

As shown in Figure III.1, the next step after model identification is parameter estimation. In Chapter 6 procedures are given to obtain efficient estimates for the parameters of a nonseasonal ARMA or ARIMA model and it is explained how an information criterion can be employed for model selection after the value of the maximized likelihood is known. Chapter 7 then deals with methods for checking the adequacy of the fitted models to ensure that relevant modelling assumptions have not been violated. Although only nonseasonal models are considered in Chapters 5 to 7, the three stage approach to model construction is also utilized for the other types of stochastic models which are discussed in this book. Furthermore, numerous practical applications demonstrate the flexibility and usefulness of the procedures which are presented.

5.2 MODELLING PHILOSOPHIES

5.2.1 Overview

Hydrologists are aware of certain types of problems which arise when modelling natural time series and these issues are outlined in Section 5.2.2. Since the practitioner is usually confronted with selecting the most suitable model from a large set of possible models for fitting to a given time series, the general topic of model discrimination is addressed in Section 5.2.3. When choosing the most appropriate model, the fundamental modelling principles of Section 5.2.4 can be satisfied by following the three stages of model building described in the general introduction to Part III as well as Section 5.2.5. Other issues related to the philosophy of model building are discussed in Section 1.3. Finally, for an overview on the philosophy of model building as well as an earlier version of the ideas expressed in this section, readers can refer to a paper presented by Hipel (1993) at a stochastic hydrology conference held in Peniscola, Spain, in 1989.

5.2.2 Hydrological Uncertainties

Engineers are concerned with the role that uncertainty plays in the design, analysis, operation and control of water resource and environmental systems. When a stochastic or time series model which is fitted to a hydrological time series is to be employed in various water resources applications, three types of uncertainties have been delineated (Kisiel and Duckstein, 1972; Wood and Rodriguez-Iturbe, 1975; Vicens et al., 1975; Wood, 1978). Firstly, there is *natural uncertainty* which is the uncertainty inherent in the natural phenomenon itself. By fitting a suitable time series model to the time series measured from the phenomenon under consideration, it is hoped that this natural uncertainty will be reflected in the mathematical structure of the model. Because the parameters of the model must be estimated statistically from a finite amount of historical data, the second kind of uncertainty is labelled *parameter uncertainty*. Finally, due to the fact that a particular model of the phenomenon may not be the "true" or best model, this creates a third category of uncertainty which is *model uncertainty*. Since the latter two types of uncertainty are dependent upon the available data, these have been jointly referred to as *information uncertainties* (Vicens et al., 1975).

Traditionally, the field of stochastic hydrology has been mainly concerned with the problem of natural uncertainty. A host of stochastic models have been developed to model natural time series and many of these models are discussed throughout this book. For instance, in addition to ARMA models, fractional Gaussian noise models and approximations to FGN have been suggested for modelling annual geophysical data sequences (see Part V). Parameter uncertainty can be measured by the standard errors for the parameter estimates and a procedure for incorporating parameter uncertainty into simulation studies is presented in Section 9.7. As reported by hydrological researchers (Vicens et al., 1975; Wood, 1978), little work has been done regarding the issue of model uncertainty. Consequently, within this chapter as well as other parts of the book, methods are described for alleviating the problem of model uncertainty. Finally, Beck (1987) provides a comprehensive review of the analysis of uncertainty in water quality modelling.

5.2.3 Model Discrimination

Model uncertainty arises because the practitioner must select the most appropriate model from the total array of models which are available for fitting to a given time series. Hence, discrimination procedures are required for choosing the most suitable model. The basic idea behind model selection is to choose a model from the set of models under consideration such that the selected model describes the data best according to some criterion. Ljung (1978) presents a unified description of *model discrimination methods* and other comprehensive articles can be found in the available literature (see for example Caines (1976, 1978) and Kashyap and Rao (1976)). Criteria for choosing the most suitable model include the capability of a model to satisfy the identification standards in Sections 5.3.2 to 5.3.7, the requirement that the model residuals pass sensitive *diagnostic checks* (see Chapter 7), the ability of a model to *forecast accurately* (see Table 1.6.3 for a summary of the forecasting work in the book), the capability of a model to *preserve important historical statistics* (see Sections 10.6 and 14.8 for nonseasonal and seasonal models, respectively), and other methods which are discussed in Section 6.3. A particularly flexible approach to model discrimination is the *Akaike information criterion* (AIC) (Akaike, 1974) which is described in Section 6.3 and initially referred to in Section 1.3.3.

5.2.4 Modelling Principles

An attractive feature of the AIC is that it automatically accounts for the certain fundamental modelling principles. As expressed by the principle of *Occam's razor* described in Section 1.3.1, one precept of stochastic model building is to keep the model as simple as possible. This can be effected by developing a model which incorporates a minimum number of parameters in order to adequately describe the data. Box and Jenkins (1976) recommend adhering to the *principle of model parsimony* (i.e. keeping the number of model parameters to a minimum) and this rule has also been of concern to hydrologists (see, for example, Jackson (1975), Tao and Delleur (1976), Hipel et al. (1977) and McLeod et al. (1977)). Besides designing a parsimonious model, a second modelling tenet is to develop a model that imparts a *good statistical fit* to the data. To achieve a good statistical fit, efficient estimates must be obtained for the model parameters (Chapter 6) and the fitted model must pass rigorous diagnostic checks to insure that the underlying modelling assumptions are satisfied (Chapter 7).

5.2.5 Model Building

In practice the key modelling doctrines of parsimony and good statistical fit can be satisfied by following the identification, estimation and diagnostic check stages of model construction. This common sense approach to model development has been advocated by both statisticians and engineers (see for example Box and Jenkins (1976), Box and Tiao (1973), Kempthorne and Folks (1971), Tao and Delleur (1976), and Hipel et al. (1977)). A flow chart for carrying out model construction is displayed in Figure III.1. As is explained in Section 6.3, an *information criterion* can be used in conjunction with the *three model building stages* in order to arrive at a simple model which fits the data well.

5.3 IDENTIFICATION METHODS

5.3.1 Introduction

When modelling a given data set a large number of models are often available for consideration. The purpose of the identification stage is to ascertain the subset of models that appear to hold more promise for adequately modelling the time series. For the case of nonseasonal ARIMA models it is necessary to determine the order of differencing if homogeneous nonstationarity is present, to ascertain the approximate number of AR and MA parameters that are required, and possibly to decide if a Box-Cox transformation is needed (see Section 3.4.5 for a discussion of the Box-Cox transformation). When the observations are stationary, differencing is, of course, not required and one must only decide upon the ARMA model parameters that are needed for adequately describing the time series that may be transformed using a Box-Cox transformation. By employing the simple graphical identification tools described in this section, usually the number of models which are worthwhile entertaining can be reduced to just a few models. In many applications, the best ARMA or ARIMA model is readily evident from the identification studies. Although each identification technique is discussed separately, in practical applications the output from all the techniques is interpreted and compared together in order to design the type of model to be estimated.

5.3.2 Background Information

Important ingredients to the identification stage are a sound understanding of the phenomenon being modelled and also an appreciation of the mathematical attributes and limitations of the stochastic models that are being considered to model the observations from that phenomenon. For example, as noted in Section 2.4.1 it is often reasonable to assume that stationary models can be fit to many kinds of annual hydrological and geophysical series of up to a few hundred years in length if the series have not been significantly influenced by external interactions. When there are external interventions such as certain types of land use changes, the effects of the interventions upon the mean level of the time series being modelled can be readily handled by employing the intervention model of Part VIII. However, when no interventions are present, it is argued in Chapter 10 that the inherent mathematical properties of the ARMA models make these models more attractive for modelling annual data than the less flexible fractional Gaussian noise models. This fact is further substantiated by using rigorous discrimination procedures to determine which type of model is more suitable according to criteria such as the Akaike information criterion (see Section 6.3) and also forecasting ability (see Chapter 8).

No matter what class of models is being entertained for modelling a given time series, the success of any modelling study is of course highly dependent upon the quantity and quality of the data (see Sections 1.2.3 and 19.7). With regard to the minimum amount of information that should be available when fitting a model to a time series various "rules of thumb" have been suggested. In a typical ARIMA modelling application of nonseasonal data it is usually preferable that there be a minimum of about 50 data points to get reasonably accurate MLE's (maximum likelihood estimates) for the model parameters (Box and Jenkins, 1976). For a fixed number of model parameters, the smaller the number of observations the larger the SE's (standard errors) of the parameter estimates will be and, hence, the relative magnitude of the SE's and parameter estimates can be examined when there are not many data points. If the SE's are quite large, the fitted model should be used with caution in certain kinds of applications and for simulation studies it may be advisable to consider parameter uncertainty as is discussed in Section 9.7. Another means to check roughly if there are sufficient data is to consider the ratio of the number of observations to the number of model parameters. If this ratio is less than three or four to one, some researchers have recommended either a more parsimonious model should be employed or else the model should not be utilized until more information becomes available. Consequently, because seasonal models (see Part VI) almost always require more parameters than nonseasonal models the minimum number of data points needed is lower for nonseasonal models.

Nonseasonal models are fitted to data such as annual riverflows and precipitation series which must be collected over quite a few years. Accordingly, for present day purposes the design of a data collection procedure may not be of immediate concern to the modeller since only the information which is currently available can be used when fitting stochastic models to observed time series. Nonetheless, the quality of data can be greatly enhanced by collecting the information properly and, consequently, network design is of great import to engineers (see Section 1.2.3). Knowing the mathematical properties of the model which may be eventually used to analyze the collected data may aid in the design of the data collection scheme. For example, Lettenmaier et al. (1978) suggest how data should be collected based upon the power of the intervention model (see the discussion in Section 19.7).

After a data collection scheme has been implemented, various factors can affect the quality of a data set. If there are errors in the measurement of the time series, this may influence the form of the model which is fitted to the data sequences. When the measurement errors are known, they should be removed before fitting a model to the time series. Systematic errors may adversely affect the estimates for the AR and MA parameters whereas random measurement errors may inflate the size of the estimated variance for the model residuals.

Often there are one or more missing values in a given time series. This is especially true for an environmental time series such as water quality measurements where data are sometimes not collected on a regular basis. When measuring riverflows, the measuring gauge may break down occasionally or perhaps may become inaccessible during severe climatic conditions and hence methods are needed to estimate the missing information. In Section 19.3, a number of useful approaches are described for estimating missing observations. For example, when there are only a few missing values, a special type of intervention model can be used.

When there is a known intervention, this can be accounted for by properly designing an intervention model (see Part VIII). For example, in Section 19.2.4, the effect of the Aswan dam upon the average annual flows of the Nile River is conveniently modelled using the method of intervention analysis. However, in certain situations the time of occurrence of an intervention or the fact that there was an intervention may not be known. For instance, the date when a precipitation gauge was replaced by a new type of gauge may not have been recorded and eventually the changing of the gauge may have been completely forgotten. Likewise, the relocation of a precipitation gauge may not have been written down in the book where the historical data are listed. Potter (1976) maintains that some precipitation time series in the United States may be "non-homogeneous" due to unknown interventions such as those just mentioned. Whatever the reason, unknown interventions sometimes occur and should be watched for when analyzing data sequences so that the series can be properly modelled.

To check for the presence of unknown interventions and also other statistical characteristics of a given time series, simple graphical procedures can be employed. Tukey (1977) refers to the numerical detective work required to discover important properties of the data as *exploratory data analysis*. A wide variety of simple graphical and numerical methods are available for use in exploratory data analysis. These methods are especially useful for dealing with messy environmental data, which may, for example, have many missing observations, be nonnormally distributed and possess outliers. A detailed discussion of exploratory data analysis is presented in Part X along with extensive water quality applications while introductory comments are put forward in Section 1.2.4. The exploratory data analysis methods described in Section 22.3 are:

1. time series plots (Section 22.3.2);
2. box-and-whisker graphs (Section 22.3.3);
3. cross-correlation function (Section 22.3.4);
4. Tukey smoothing (Section 23.3.5);
5. autocorrelation function (Section 23.3.6 and [2.5.4]).

In this section, exploratory techniques which are specifically well designed for identifying ARMA or ARIMA models are discussed. Section 5.3.3 and also Section 22.3.2 explain how a plot of the time series under consideration can reveal many of the essential mathematical features of the data. After surveying the general properties of the series using a plot of the series or other exploratory data analysis tools, the identification techniques described in Sections 5.3.4 to 5.3.7 are employed for determining the approximate orders of the operators of an ARMA or ARIMA model which could be fitted to the time series.

After a model has been fitted to the sample data, *confirmatory data analysis* techniques can be employed to investigate the capabilities or characteristics of the fitted model (Tukey, 1977) and, hence, the data set it describes. For example, in Section 10.6 it is shown how significance testing can be used to determine whether or not important historical statistics are preserved by the fitted model. In Section 8.3 the relative forecasting performance of the different kinds of nonseasonal models are examined. A general description of both exploratory and confirmatory data analysis is presented in Sections 1.2.4 and 22.1 as well as the overview to Part X.

5.3.3 Plot of the Data

A visual inspection of a *graph of the given observations* against time can often reveal both obvious and also less apparent statistical characteristics of the data. Identification information which may be gleaned from a perusal of a graph include:

1) *Autocorrelation* - Linear dependence existing among observations may cause certain types of loose patterns in the data. For instance, at certain sections of the time series the observations may be consistently above an overall mean level whereas at other locations values below the mean level may be grouped together. Hydrologists refer to this property as *persistence* and from a statistical viewpoint this means that the data are probably autocorrelated. The form of the data set displayed in Figure 2.3.1 shows that the historical observations of the annual flows of the St. Lawrence River at Ogdensburg, New York are correlated. The same conclusion holds for the simulation sequences in Figures 2.3.2 and 2.3.3 which were generated by the AR model in [3.2.19] fitted to the St. Lawrence flows. A situation where the data do not seem to follow any kind of pattern may indicate that the time series is white noise. The behaviour of a white noise sequence is exemplified by the simulated white noise series in Figure 4.3.2.

2) *Seasonality* - Usually it is known in advance whether or not a data set is seasonal and a graphical display will simply confirm what is already obvious. For geophysical data seasonality is of course caused by the annual rotation of the earth about the sun and hence usually only annual data are nonseasonal. Figure VI.1 at the start of the part of the book on seasonal models displays a graph of the average monthly flows in m^3/s of the Saugeen River at Walkerton, Ontario, Canada, from January 1915 until December 1976. The cyclic behaviour of the graph demonstrates that the series is indeed seasonal. Three types of seasonal models for fitting to time series are described in Chapters 12 to 14 in Part VI.

   In certain situations it may not be obvious before examining a plot of the data whether or not a given series is seasonal. This may be the case for a socio-economic time series such as monthly water demand for a highly industrialized city located in a temperate climate. Some types of monthly or weekly pollution time series may also exhibit nonseasonality. For example, in Section 19.4.5 a nonseasonal intervention model is fitted to the series of monthly phosphorous levels in a river shown in Figure 1.1.1.

3) *Nonstationarity* - The presence of nonstationarity is usually suspected or known before plotting the time series. The explosive type of nonstationarity which is discussed in Section 4.2 may be indicated by plots similar to those given in Figures 4.2.1 and 4.2.2.

Examples of homogeneous nonstationarity described in Section 4.3 are shown in Figure 4.3.1 and also Figures 4.3.3 to 4.3.5. Other illustrations of homogeneous nonstationarity are displayed by the annual water use series in Figure 4.3.8, the yearly electricity consumption in Figure 4.3.10, and also the Beveridge wheat price indices in Figure 4.3.15. These figures clearly indicate various manners in which data may not follow an overall mean level.

4) *Trends* - The presence of trends in the data is a form of nonstationarity. As discussed in Section 4.6, trends can be classed as either deterministic or stochastic. Deterministic trends can be expressed as a function of time as shown in [4.6.1] whereas stochastic trends can often be accounted for by using the differencing operator of sufficiently high order in [4.3.3]. If trends are present in the plot of a data that do not appear to follow the path of a deterministic function but rather evolve in a stochastic fashion, then differencing may account for these trends. Trends may not only affect the level of a series but they may also be associated with changes in variance in the series. Consider, for example, the average monthly water useage in millions of litres per day depicted in Figure VI.2 for the city of London, Ontario, Canada, from January, 1966, until the end of December, 1988. This figure reveals that the water demand data fluctuates in a cyclic pattern due to the seasonality and contains a linear trend component coupled with an increase in variance in later years as the data spreads further apart around the linear trend for increasing time. An appropriate Box-Cox transformation from [3.4.30] has the effect of pulling the data together and reducing the change in variance over time for the time series given in Figure VI.2 and modelled in Section 12.4.2.

5) *Need for a transformation* - Figure VI.2 is an example of a data plot where it appears from a graph of the original data that a *Box-Cox transformation* is needed. If a transformation is required but this fact is not discovered at the identification stage, the need for a data transformation will probably be detected at the diagnostic check stage of model development when the properties of the residuals are examined (see Chapter 7). In practice, it has been found that a transformation of the data usually does not affect the form of the model to fit to the data (i.e. the orders of p, d and q in an ARIMA(p,d,q) model). However, this is not true for all situations and as pointed out by Granger and Newbold (1976), certain transformations can change the type of model to estimate. Consequently, when a specific form of transformation is decided upon at the identification stage, it is preferable to complete all three stages of model construction using the transformed data. On the other hand, if the requirement for a Box-Cox transformation is not determined until the diagnostic check stage, it is usually not necessary to repeat the identification stage for the transformed data. Instead, the parameters of the model can be estimated for the transformed data and only if diagnostic checks reveal the model is unsatisfactory would it be necessary to return to the identification stage to ascertain the proper orders of p, d and q.

6) *Extreme values* - The presence of extreme values or outliers is easily detected in a graphical display of the data. When dealing with riverflow time series, large values could be due to excessive precipitation while extremely low flows occur during times of drought. If investigation into the collection and processing of the data indicates that the extreme values appear to be correct, various courses of action are available to ensure that the outliers are properly handled. When an outlier is caused by a known external intervention, an intervention component can be introduced into the model to allow for this (see Chapter 19).

Sometimes a transformation such as a Box-Cox transformation (see Section 3.4.5) may reduce undesirable consequences that outliers may have in stochastic model building. For example, taking natural logarithms of the data may pull the observations together so that the outliers do not have a significant detrimental effect upon the residuals of the fitted model. Other types of data transformations are also discussed by Granger and Orr (1972). Of particular interest is the method of *data clipping* which was also used for various types of applications by Tukey (1962), Rothenberg et al. (1964), Fama and Roll (1968, 1971) and Rosenfeld (1976). To clip the time series, the data are firstly ranked from smallest to largest. If it is desired to clip only the larger observations, the last k percent of values are replaced by the mean of the remaining $(100-k)$ percent of data. When it is required to clip both the smaller and larger outliers, the last $k/2$ percent and also the first $k/2$ percent of values can be removed and then replaced by the mean of the remaining $(100-k)$ percent of data. If the clipped and unclipped time series produce similar results at the three stages of model construction, then the outliers do not hinder the stochastic model building procedure. However, if the results differ, appropriate action may be taken. For instance, after transforming the data using a Box-Cox transformation, the models which are selected to fit to both the clipped and unclipped data of the transformed time series, may be the same. Rosenfeld (1976) discusses the use of data clipping in model identification. If, for example, an important identification feature such as a large value of the sample ACF at a given lag appears for both the clipped and unclipped data, it is likely to be a true feature of the model. On the other hand, Rosenfeld (1976) claims that if a significant identification characteristic in the original time series is lost by clipping, it is probably the result of coincidentally placed extreme outliers. In situations where clipping results in an identifying feature which does not appear in the original identifying function such as the sample ACF, it is probably caused by the clipping and is not a true feature of an underlying model.

7) *Long term cycles* - Often natural data sets are too short to detect any long term cycles which, for instance, may be due to gradual changes in climate. However, tree ring index series are available for time spans of thousands of years (Stokes et al., 1973) and hence for certain data sets it may be possible to graphically detect long term cycles.

8) *Known or unknown interventions* - The effects of a known intervention can often be detected by an examination of the plot of the time series. For example, Figure 19.2.1 clearly portrays the drop in the mean level of the annual flow of the Nile River at Aswan, Egypt, due to the construction of the Aswan dam in 1902. The yearly flows are calculated for the water year from October 1st to September 30th of the following year and are available from October 1, 1870 to October 1, 1945. An intervention model for the Nile River data is designed in Section 19.2.4.

   When a data plot indicates that there may be a significant change in the mean level due to an unknown intervention, an investigation should be carried out to see if a physical reason can be found. For instance, as discussed in Section 5.3.2 precipitation records may be significantly affected by changing the type gauge. If it is ascertained that there is a physical cause for the mean level change, an intervention model can be developed (see Part VIII). Alternatively, the apparent change in the mean of the series may only be due to inherent natural fluctuations in the series and a regular ARMA model may adequately model the data.

5.3.4 Sample Autocorrelation Function

By utilizing [2.5.9], the *sample autocorrelation function* (ACF) of a time series can be calculated and then plotted against lag k up to a maximum lag of approximately $N/4$ where N is the length of the series. If the theoretical ACF is assumed to be zero after lag q, [2.5.11] can be used to calculate confidence limits. When it is not certain beyond which lag ρ_k is zero, it is often convenient to start out by plotting the confidence limits for white noise (i.e. ρ_k is assumed to be zero after lag zero).

As noted in Section 5.3.3, it is often known in advance whether or not the series under consideration is nonstationarity. A plot of the series will usually reveal nonstationarity, although when the data are only marginally nonstationary it may not be certain as to whether differencing is required to account for homogeneous nonstationarity. In Section 4.3.2 it was explained why the ACF of a process which possesses homogeneous nonstationarity attenuates slowly. Consequently, when the sample ACF of the given nonseasonal data set dies off slowly it may be advisable to difference the data once. If the sample ACF of the differenced series still does not damp out quickly, the series should be differenced again. The data should be differenced just enough times to remove the homogeneous nonstationarity which in turn will cause the sample ACF to die off rather quickly. When differencing is required, usually it is not greater than 2 for nonstationary series which arise in practice.

Following differencing, the resulting stationary w_t series of length $n = N - d$ in [4.3.3] is examined to determine the orders of p and q. If the given data is approximately stationary, then the w_t series is in fact the $z_t^{(\lambda)}$ data set and, hence, the properties of the $z_t^{(\lambda)}$ series are investigated to determine how many AR and MA terms may be needed in the model. When the sample ACF of the stationary w_t series is plotted along with the appropriate confidence limits up to a lag of about $n/4$, the following general rules may be invoked to help to determine the orders of p and q.

1) If the series can be modelled by a white noise model, then r_k in [2.5.9] is not significantly different from zero after lag zero. From Section 2.5.4, r_k is approximately NID(0,1/n).

2) For a pure MA model, r_k cuts off and is not significantly different from zero after lag q.

3) When r_k damps out and does not appear to truncate, this suggests that AR terms are needed to model the time series.

5.3.5 Sample Partial Autocorrelation Function

The theoretical definition for the partial autocorrelation function (PACF) is given by the Yule-Walker equations in [3.2.17] while the algorithm of Pagano (1972) for estimating the values of the PACF is outlined in Appendix A3.1. Assuming that the process is AR(p), the estimated values of the PACF at lags greater than p are asymptotically normally independently distributed with a mean of zero and standard error of $1/\sqrt{n}$ in [3.2.18]. Because the asymptotic distribution is known, one can plot 95% confidence limits. When differencing is required, the sample PACF is only plotted for the stationary w_t series in [4.3.3] up to a lag of about $n/4$.

When used in conjunction with an identification aid such as a plot of the sample ACF, the estimated PACF is useful for determining the number of AR and MA parameters. The following general characteristics of the PACF may be of assistance in model identification.

1) When the series is white noise, the estimated values of the PACF are not significantly different from zero for all lags.

2) For a pure AR model, the sample PACF truncates and is not significantly different from zero after lag p.

3) If the sample PACF attenuates and does not appear to cut off, this may indicate that MA parameters are needed in the model.

5.3.6 Sample Inverse Autocorrelation Function

Cleveland (1972) defines the *inverse autocorrelation function (IACF)* of a time series as the ACF associated with the reciprocal of the spectral density function of the series. The theoretical IACF, ξi_k can also be specified in an alternative equivalent fashion within the time domain. When considering the ARIMA(p,d,q) process in [4.3.4], the theoretical IACF of w_t in [4.3.3] is defined to be the ACF of the (q,d,p) process which is written as

$$\theta(B)w_t = \phi(B)a_t \qquad [5.3.1]$$

A similar definition for the theoretical IACF also holds for the seasonal case. The theoretical IACF is the ACF of the process where not only the nonseasonal AR and MA operators have been interchanged but the seasonal AR and MA operators have also been switched (see Section 12.3.2 for a description of identification tools for seasonal ARIMA models).

Besides the original paper of Cleveland (1972), applications and theoretical developments regarding the IACF are given in papers by Hipel et al. (1977), McLeod et al. (1977), Chatfield (1979), Hosking (1980), Bhansali (1980, 1983a,b), Abraham and Ledolter (1984), and Battaglia (1988). The IACF is also mentioned briefly by Parzen (1974), McClave (1975, p. 213), Granger and Newbold (1977, p. 109) and also Shaman (1975). As noted by Cleveland (1972), one reason why the IACF was not a popular identification tool may be due to the fact that the reciprocal of the spectrum is not an intuitively meaningful quantity. Certainly, the time domain definition of the theoretical IACF which is employed by Hipel et al. (1977) and Chatfield (1979) is much more appealing.

Another explanation why the IACF was not used extensively in the past may be caused by the lack of a good estimation procedure for determining the sample IACF for a given time series (Hipel et al., 1977, p. 569). However, progress has been made on developing estimation techniques for calculating the sample IACF (Bhansali, 1983a,b; Battaglia, 1988). To obtain an estimate ri_k for ξi_k at lag k, Cleveland (1972) suggests using either an AR or smoothed periodogram estimation procedure. If the AR approach is adopted, the first step is to model the w_t series by an AR model of order r. The estimates $\hat{\phi}_i$ where $i = 1,2, \ldots, r$, for the AR parameters, can be determined from the Yule-Walker equations in [3.2.12] or from the maximum likelihood estimates of an AR(r) model which is fit to the time series under consideration. The estimate ri_k for the theoretical IACF at lag k can then be obtained from

$$ri_k = \left[-\hat{\phi}_k + \sum_{i=1}^{r-k}\hat{\phi}_i\hat{\phi}_{i+k}\right]\left[1 + \sum_{i=1}^{r}\hat{\phi}_i^2\right]^{-1} \qquad [5.3.2]$$

If the w_t series is white noise, ri_k is approximately NID(0,1/n).

To utilize the IACF for model identification calculate and plot ri_k versus lag k, where ri_k can go from -1 to +1. A recommended procedure is to choose about four values of r between 10 and 40 (where $r < n/4$) and then to select the most representative graph from the set for use in identification. One of the reasons why Hipel et al. (1977) suggest that an improved estimation procedure should be developed for the IACF is because a selection procedure is needed to choose an appropriate plot of the sample IACF. From a knowledge of the distribution of ri_k, confidence limits can be drawn on the graph of the sample IACF. For white noise, ri_k is approximately NID(0,1/n) while for a correlated series ri_k is normally distributed with a mean of zero and the variance of ri_k after lag p is given by

$$var(ri_k) \approx \frac{1}{n}\left\{1 + 2\sum_{j=1}^{p} ri_j^2\right\}, \; k > p \qquad [5.3.3]$$

When using the sample IACF for model identification to ascertain the orders of p and q, the following rules may be utilized:

1) If the series can be modelled by a white noise model, ri_k is not significantly different from zero after lag zero.

2) For a pure AR model, ri_k truncates and is not significantly different from zero after lag p. In practice, it has been found that the IACF is useful for identifying AR models where some of the AR parameters should be constrained to be zero (see Section 3.4.4 for a discussion of constrained models). At the same lags at which the AR parameters are zero, the sample IACF often possesses values that are not significantly different from zero (Cleveland, 1972; Hipel et al., 1977; McLeod et al., 1977).

3) When ri_k attenuates and does not appear to cut off, this indicates that MA terms are needed to model the time series.

As can be seen, the foregoing general properties of the sample IACF are similar to those listed for the sample PACF in Section 5.3.5. Due to this fact and also the estimation problems with the IACF, the sample IACF is not used extensively by practitioners. However, as shown by the applications in Section 5.4, the sample IACF along with other identification graphs are very helpful when employed together for identifying ARMA models. Furthermore, Cleveland (1972) has shown how the sample IACF can be utilized for identifying the components in transfer function-noise models (see Part VII for a presentation of these models). In fact, Cleveland (1972) recommends using the sample ACF and IACF for model identification rather than the sample ACF and PACF.

5.3.7 Sample Inverse Partial Autocorrelation Function

Hipel et al. (1977) provide the original definition of the *inverse partial autocorrelation function (IPACF)* as the PACF of an ARMA(q,p) process. To define mathematically the theoretical IPACF, consider the inverse Yule-Walker equations given by

$$\begin{bmatrix} 1 & \rho i_1 & \rho i_2 & \cdots & \rho i_{k-1} \\ \rho i_1 & 1 & \rho i_1 & & \rho i_{k-2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \rho i_{k-1} & \rho i_{k-2} & \rho i_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \theta_{k1} \\ \theta_{k2} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{kk} \end{bmatrix} = \begin{bmatrix} \rho i_1 \\ \rho i_2 \\ \cdot \\ \cdot \\ \cdot \\ \rho i_k \end{bmatrix} \qquad [5.3.4]$$

where ρi_k is the theoretical IACF at lag k and θ_{kj} is the jth coefficient in a MA process of order k such that θ_{kk} is the last coefficient.

The coefficient θ_{kk} is called the theoretical IPACF. To obtain an estimate $\hat{\theta}_{kk}$ for θ_{kk} replace ρi_k by the sample IACF ri_k and solve the inverse Yule-Walker equations for $\hat{\theta}_{kk}$. Because of the problems encountered when estimating ρi_k, another approach would be to first estimate θ_{kk} as the kth coefficient in a MA model of order k. Based upon results in the inverse Yule-Walker equations, appropriate methods could then be used to estimate ρi_k. Bhansali (1983c) presents another procedure for estimating the IPACF.

For model identification, plot $\hat{\theta}_{kk}$ against lag k for the same number of lags as were chosen for the sample IACF. The values of the sample IPACF can range from -1 to +1. Furthermore, when the theoretical IPACF is known to be zero after lag q, the sample IPACF is approximately NID(0,1/n) after lag q (McLeod, 1984). Consequently, the 95% confidence limits can be plotted on the graph of the sample IPACF. When employing the plot of the sample IPACF for model identification, the following properties can be kept in mind.

1) When the time series is white noise, the sample IPACF is not significantly different from zero after lag zero.

2) For a pure MA model, $\hat{\theta}_{kk}$ cuts off and is not significantly different from zero after lag q.

3) When the sample IPACF damps out and does not appear to truncate, this suggests that AR terms are needed in order to suitably model the series.

The inherent characteristics of the sample IPACF are similar to those of the sample ACF in Section 5.3.4. Even though this pair of functions possesses the same general properties for identifying an ARMA model to fit to a series, the two functions are defined differently. In a given situation, for instance, one identification function may more clearly reveal a characteristic of the data than the other. Consequently, both the IPACF and ACF are recommended for application to the series under consideration. Likewise, as noted earlier, common relationships also exist between the PACF and IACF, and both of these functions should also be used in the application. The general attributes of all these useful identification functions are summarized in Table 5.3.1.

Table 5.3.1. Properties of four identification methods.

Identification Method	Types of Models		
	AR(p)	MA(q)	ARMA(p,q)
ACF	Attenuates	Truncates after lag q	Attenuates
PACF	Truncates after lag p	Attenuates	Attenuates
IACF	Truncates after lag p	Attenuates	Attenuates
IPACF	Attenuates	Truncates after lag q	Attenuates

5.4 APPLICATIONS

5.4.1 Introduction

After examining a plot of a time series to pick out basic statistical properties of the data set, the sample ACF, PACF, IACF and IPACF, described in Sections 5.3.4 to 5.3.7, respectively, can be used to identify the AR and MA parameters needed in an ARMA or ARIMA model to fit to the series. Table 5.3.1 describes the main characteristics to look for when using these functions for model identification.

In Chapters 2 to 4, a variety of nonseasonal time series are examined for explaining concepts presented in those chapters. Tables 5.4.1 and 5.4.2 summarize the identification results for the stationary and nonstationary series, respectively. Notice that wherever an identification graph for a series appears in the book, the figure number is given in the tables. For illustration purposes, the manner in which the ARMA models are identified for the annual St. Lawrence riverflows and also the Wolfer sunspot numbers in Table 5.4.1 are explained in detail in this section following the modelling of these two series carried out by McLeod et al. (1977).

In Section 4.3.3, nonstationary ARIMA(p,d,q) models are identified for fitting to the following three series:

1. annual water use for New York City,
2. annual electricity consumption in the U.S.,
3. Beveridge wheat price index.

As explained in that section, sometimes a graph of the original series indicates whether or not a data transformation is needed. The next step is to ascertain the order of differencing that is required. The need for differencing can be determined from a graph of the series or the fact that the sample ACF dies off very slowly. After the data are differenced just enough times to remove nonstationarity, the sample ACF, PACF, IACF and IPACF are employed to determine the AR and MA parameters required in the ARMA(p,q) model in [4.3.4] to fit to the w_t series in [4.3.3]. Table 5.4.2 summarizes the identification results for the three aforementioned nonstationary series.

Table 5.4.1. Identification of ARMA models to fit to nonseasonal stationary series.

Annual Series (Source)	Time Series Plot	Box-Cox λ	Sample ACF	Sample PACF	Sample IACF	Sample IPACF	ARMA Model Identified
St. Lawrence flows at Ogdensburg, New York. 1860-1957. (Yevjevich, 1963)	Fig.'s 2.3.1 and 5.4.1.	1	Fig.'s 3.2.1 and 5.4.2. Dies off.	Fig.'s 3.2.4 and 5.4.3. Cuts off. Large values at lags 1 and 3.	Fig. 5.4.4. Truncates. Large values at lags 1 and 3.	Fig. 5.4.5. May die off.	Constrained AR(3) model without ϕ_2.
Temperatures from English Midlands 1813-1912. (Manley, 1953)		1	Fig.'s 2.5.1 and 2.5.2. Truncates. Large values at first two lags.	Fig. 3.3.1. Cuts off. Large values at lags 1 and 2.	Truncates. Big values at first two lags.	Cuts off. Large values at first two lags.	AR(2) or MA(2).
Rhine River flows at Basle, Switzerland 1837-1957. (Yevjevich, 1963)		1	Fig. 2.5.4. White noise.	White noise.	White noise.	White noise.	ARMA(0,0)
Douglas fir tree ring data at Navajo Natoinal Monument in Arizona. 1263-1962. (Stokes et al., 1973)		1	Fig. 3.4.1. Dies off.	Fig. 3.4.2. Perhaps attenuates.	Dies off.	Attenuates	ARMA(1,1)
Wolfer sunspot numbers. 1700-1869. (Waldmeier, 1961)	Fig. 5.4.6	0.5	Fig. 5.4.7 for λ=1. Cyclic. Dies off slowly.	Fig. 5.4.8 for λ=1. Large values at lags 1, 2 and around lag 8.	Fig. 5.4.9 for λ=1. Large value at lag 1.	Fig. 5.4.10 for λ=1. Large values at low lags and also around lag 2.	Constrained AR(9) model without ϕ_3 to ϕ_8.

5.4.2 Yearly St. Lawrence Riverflows

Average annual riverflows for the St. Lawrence River at Ogdensburg, New York, are available from 1860 to 1957 (Yevjevich, 1963) and are plotted in Figure 2.3.1. For convenience, these flows are also displayed in this section in Figure 5.4.1. The sample ACF and PACF, and their accompanying 95% confidence limits, for the St. Lawrence riverflows are displayed in Figures 3.2.1 and 3.2.4, respectively, as well as Figures 5.4.2 and 5.4.3, respectively, in this section. In addition, the sample IACF and IPACF, along with the 95% confidence limits, for the St. Lawrence flows are drawn in Figures 5.4.4 and 5.4.5, respectively. In practice, one can quickly peruse these graphs as they are displayed on a computer screen in order to identify the type of ARMA model to fit to the series.

Because the St. Lawrence riverflows appear to fluctuate around an overall mean level and not follow a trend in Figure 5.4.1, one can argue that the flows are stationary. This fact is also confirmed by the behaviour of the sample ACF in Figure 5.4.2, which dies off fairly quickly. Since the sample ACF does not truncate but rather damps out, this suggests that AR parameters are needed in the ARMA model to fit to the series. The 95% confidence limits for the graph of the sample PACF in Figure 5.4.3 are for values of the PACF at lags greater than p if the process were ARMA(p,0). Notice that the sample PACF possesses a significantly large value at lag 1

Table 5.4.2. Identification of ARIMA models to fit to nonseasonal nonstationary series (see Section 4.3.3).

Annual Series (Source)	Time Series Plot	Box-Cox λ	Sample ACF	Sample PACF	Sample IACF	Sample IPACF	ARIMA Model Identified
Water use for New York City. 1898-1968. (Salas and Yevjevich, 1972)	Fig. 4.3.8	1	Because sample ACF in Fig. 4.3.9 dies off slowly, differencing is needed. Differenced series is white noise.	Differenced series is white noise.	Differenced series is white noise.	Differenced series is white noise.	ARIMA(0,1,0)
Electricity consumption in U.S. 1920-1970. (United States Bureau of the Census, 1976)	Fig. 4.3.10	0.533	Sample ACF in Fig. 4.3.11 for given series and also sample ACF in Fig. 4.3.12 for differenced series die off slowly. Hence, order of differencing needed is $d=2$. Sample ACF in Fig. 4.3.13 for data differenced twice has large value at lag 1.	Sample PACF for $d=2$ in Fig. 4.3.14 can be interpreted as attenuating.	Sample IACF for series with $d=2$ dies off.	Sample IPACF for series with $d=2$ cuts off after lag 1.	ARIMA(0,2,1)
Beveridge Wheat Price Index. 1500-1869. (Beveridge, 1921)	Fig. 4.3.15 shows level and variance are increasing with time.	0	Since sample ACF in Fig. 4.3.16 for logarithmic data attenuates slowly differencing is needed. Sample ACF in Fig. 4.3.17 for differenced logarithmic data has large values at low lags and lag 8.	Sample PACF in Fig. 4.3.18 for logarithmic data with $d=2$ has large values at lags 2 and 8.	Sample IACF for logarithmic data with $d=2$ has large values at low lags and lag 8.	Sample IPACF for logarithmic data with $d=2$ has large values at low lags and lag 8.	Constrained ARIMA(8,1,1) model without ϕ_3 to ϕ_7

and has a value at lag 3 that just touches the upper 95% confidence limit. This effect is more clearly illustrated by the sample IACF in Figure 5.4.4 which has definite large values at lags 1 and 3. It may, therefore, be appropriate to entertain an ARMA(3,0) model with ϕ_2 constrained to zero as a possible process to fit to the St. Lawrence River data. Although there are rather large values of the estimated PACF at lag 19 and of the sample IACF at lag 18, these could be due to chance alone. The sample IPACF in Figure 5.4.5 appears to be attenuating rather than truncating. However, for this particular example the sample ACF definitely damps out, and therefore one would suspect that the sample IPACF is behaving likewise, thereby indicating the need for AR terms. On the graph for the sample IPACF, the 95% confidence intervals are for values of the IPACF at lags greater than q if the process were ARMA(0,q).

For the case of the Saint Lawrence River data, the sample IACF in Figure 5.4.4 most vividly defines the type of model to estimate. However, the remaining three identification graphs reinforce the conclusions drawn from the IACF.

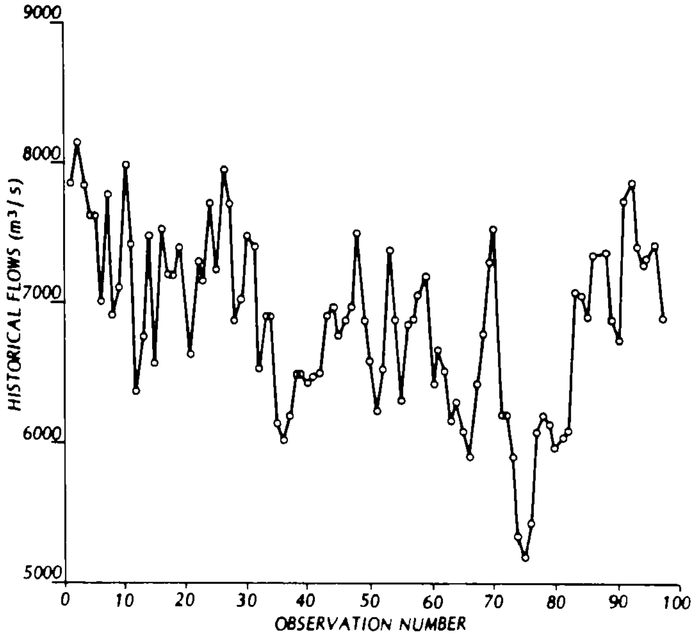


Figure 5.4.1. Annual flows of the St. Lawrence River at Ogdensburg, New York from 1860 to 1957.

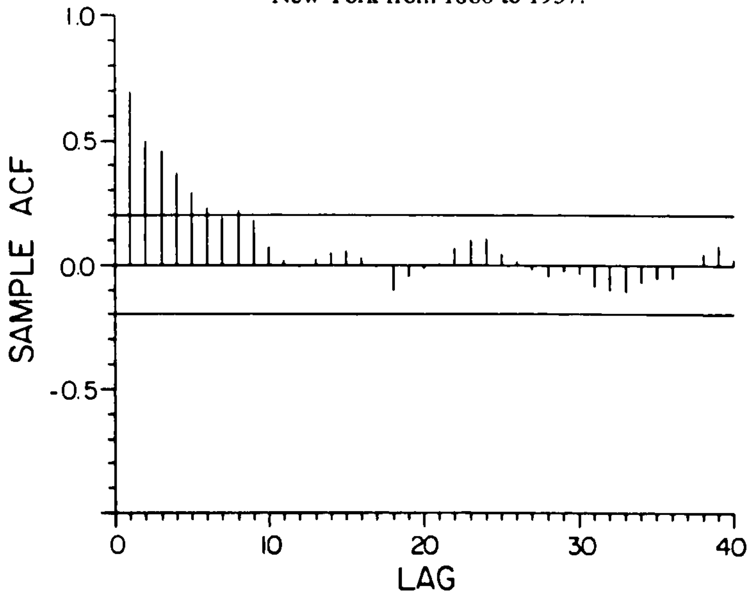


Figure 5.4.2. Sample ACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

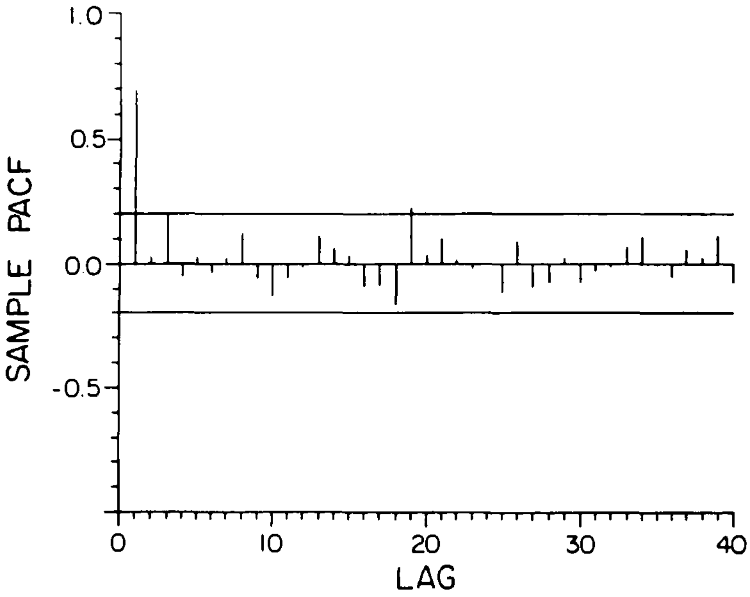


Figure 5.4.3. Sample PACF and 95% confidence limits for the average yearly flows of the St. Lawrence River at Ogdensburg, New York.

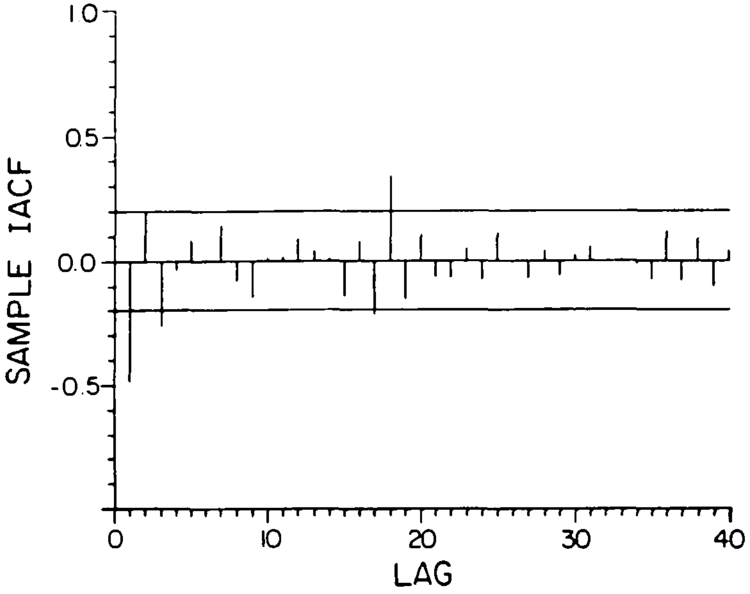


Figure 5.4.4. Sample IACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

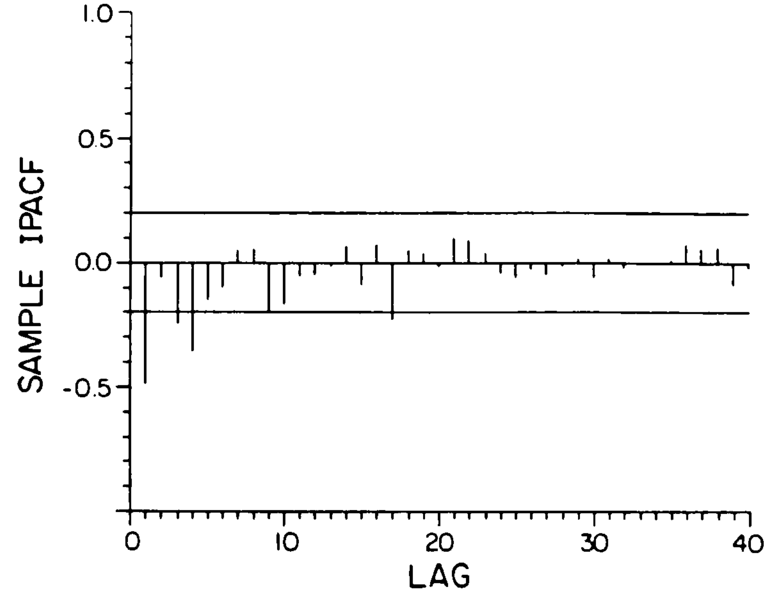


Figure 5.4.5. Sample IPACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

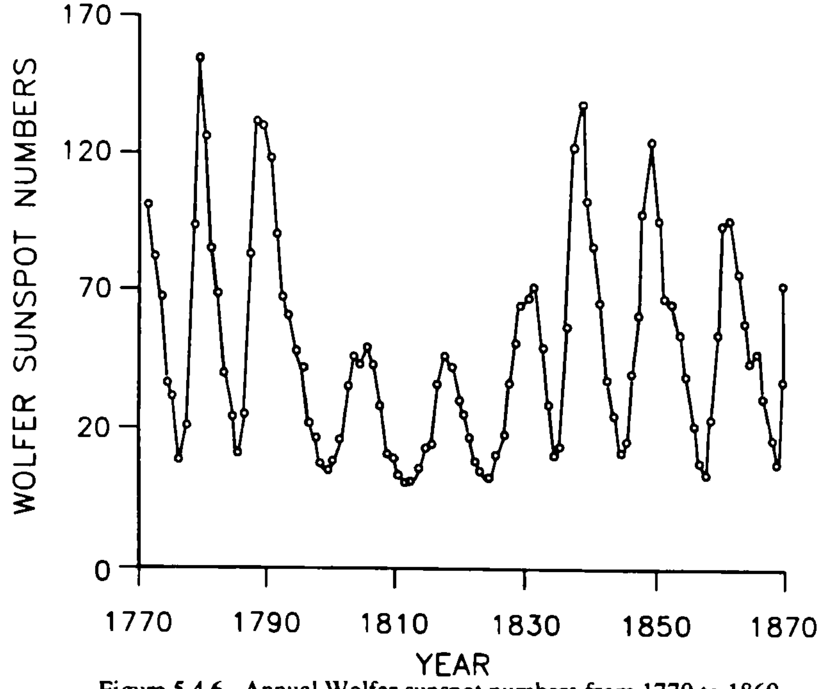


Figure 5.4.6. Annual Wolfer sunspot numbers from 1770 to 1869.

5.4.3 Annual Sunspot Numbers

Annual sunspot numbers are examined here because of the historical controversies regarding the selection of a suitable model to fit to yearly sunspot numbers and also because sunspot data are of practical importance to geophysicists and environmental scientists. Climatologists have discovered that sunspot activity may be important for studying climatic change because of its effect upon global temperature variations (Schneider and Mass, 1975). However, in Chapter 16, it is shown statistically that sunspot numbers do not affect the yearly flows of the Volga River in the Soviet Union. Nonetheless, sunspots have long been known to affect the transmission of electromagnetic signals.

The Wolfer sunspot number series is available from 1700 to 1960 in the work of Waldmeier (1961), while Box and Jenkins (1976) list the average annual sunspot series from 1770 to 1869 as series E in their textbook. Granger (1957) found that the periodicity of sunspot data follows a uniform distribution with a mean of about 11 years, and for this and other reasons researchers have had difficulties in modeling sunspot numbers. Indeed, the graph of the annual sunspot numbers from 1770 to 1869 displayed in Figure 5.4.6 clearly shows this periodicity. Yule (1927) employed an AR model of order 2 to model yearly sunspot numbers. Moran (1954) examined various types of models for predicting annual sunspot numbers and expressed the need for a better model than an ARMA(2,0) model. Box and Jenkins (1976) fitted ARMA(2,0) and ARMA(3,0) models to yearly sunspot data, Bailey (1965) entertained an ARMA(6,0) model, Davis (1979) employed ARMA(2,0) and ARMA(9,0) models, and Craddock (1967) and Morris (1977) considered AR models up to lag 30 for forecasting annual sunspot numbers. Phadke and Wu (1974) modelled yearly sunspot numbers using an ARMA(1,1) model while Woodward and Gray (1978) utilized an ARMA(8,1) model.

Other researchers have determined stochastic sunspot models when the basic time interval is smaller than one year. For example, Whittle (1954) considered a unit of time of six months and developed a bivariate AR scheme to fit to the observed sunspot intensities in the northern and southern solar hemispheres. Granger (1957) proposed a special two-parameter curve for the monthly sunspot numbers, but unfortunately, this curve is not useful for forecasting.

Even though the annual sunspot numbers are difficult to model, the identification graphs defined in Sections 5.3.4 to 5.3.7 can be used to design a reasonable ARMA model to fit to the annual sunspot series. In Section 6.3, it is explained how the Akaike information criterion (Akaike, 1974) can be used in conjunction with these graphs to come up with the same model. The sample ACF, PACF, IACF and IPACF graphs, along with their 95% confidence limits, are displayed in Figures 5.4.7 to 5.4.10, respectively, for the annual sunspot numbers. As can be seen in Figure 5.4.7, the sample ACF follows an attenuating sine wave pattern that reflects the random periodicity of the data and possibly indicates the need for nonseasonal and/or seasonal AR terms in the model. The behaviour of the sample PACF shown in Figure 5.4.8 could also signify the need for some type of AR model. In addition to possessing significant values at lags 1 and 2, the PACF also has rather large values at lags 6 to 9. The sample IACF in Figure 5.4.9 has a large magnitude at lag 1, which suggests the importance of a nonseasonal AR lag 1 term in any eventual process that is chosen to estimate. The dying out effect in the first four lags of the sample IPACF displayed in Figure 5.4.10 could be a result of a nonseasonal AR component.

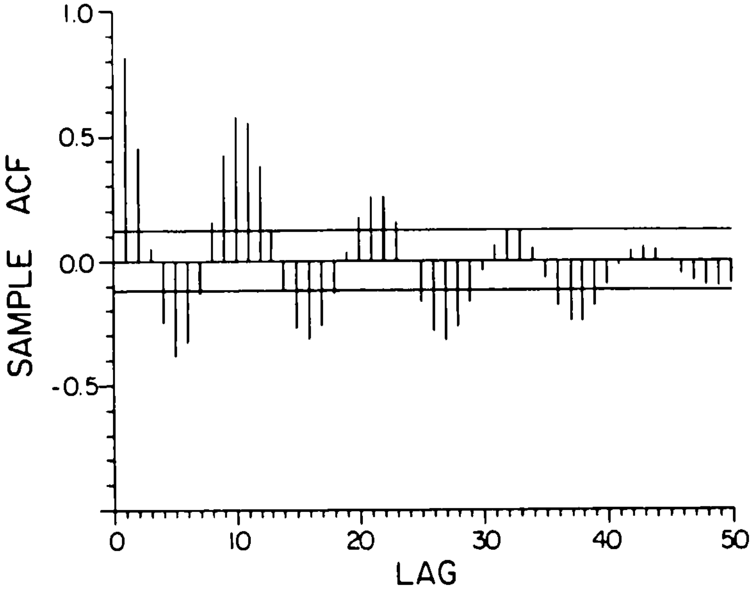


Figure 5.4.7. Sample ACF and 95% confidence limits for the yearly sunspot numbers.

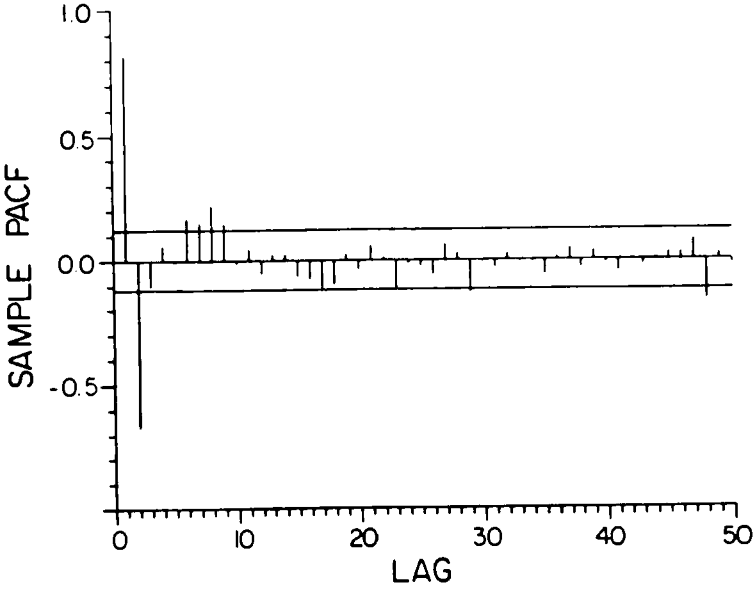


Figure 5.4.8. Sample PACF and 95% confidence limits for the annual sunspot numbers.

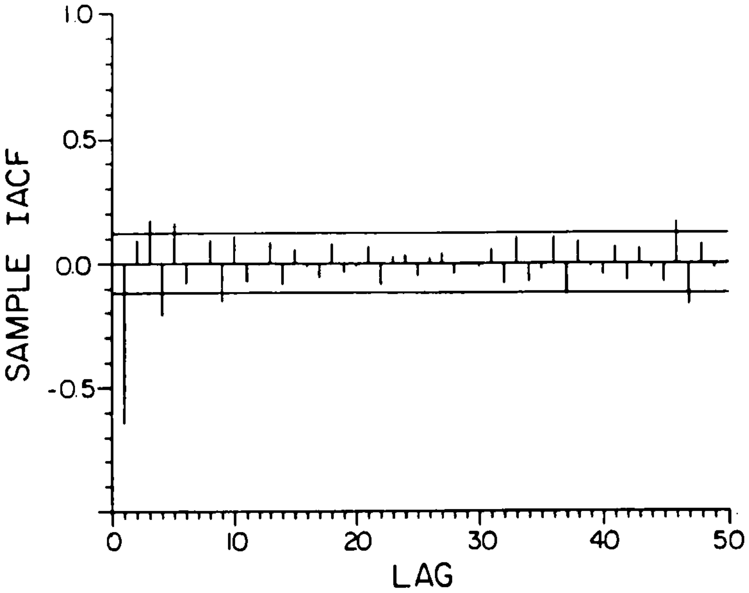


Figure 5.4.9. Sample IACF and 95% confidence limits for the yearly sunspot numbers.

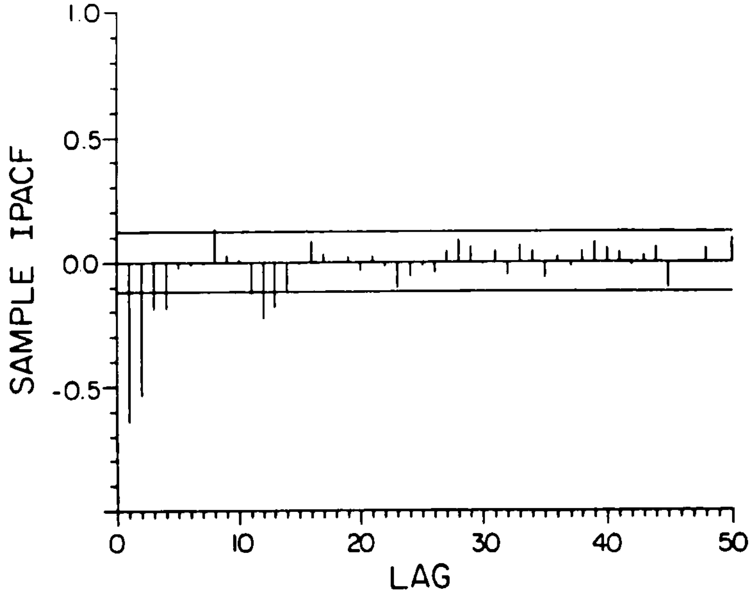


Figure 5.4.10. Sample IPACF and 95% confidence limits for the annual sunspot numbers.

Diagnostic checks are discussed in detail in Chapter 7. For the case of the sunspot numbers, results of diagnostic checks from ARMA models fitted to the series, as well as the identification graphs, are needed to come up with the best overall model. When an AR(2) model is fit to the yearly sunspot numbers, the independence, normality, and homoscedasticity assumptions of the residuals are not satisfied. As explained in Section 3.4.5, to overcome problems with nonnormality and heteroscedasticity (i.e. changing variance) in the model residuals, one can employ the Box-Cox transformation in [3.4.30]. By substituting $\lambda = 0.5$ and $c = 1.0$ in [3.4.30], one can obtain a square root transformation for the annual sunspot series. It is necessary to set $c = 1$ because there are some zero entries in the sunspot series and the Box-Cox transformation in [3.4.30] can only be used with positive values. This square root transformation causes the residuals of an AR(2) model fitted to the transformed sunspot series to become approximately normally distributed and homoscedastic. However, because the residuals are autocorrelated a better model is required to fit to the transformed series.

If an AR(3) model with $\lambda = 0.5$ and $c = 1.0$ is estimated, the $\hat{\phi}_3$ parameter has a magnitude of -0.103 and a SE of 0.062. Because $\hat{\phi}_3$ is less than twice its standard error, for the sake of model parsimony it should not be incorporated into the model. Note that Box and Jenkins (1976, p. 239, Table 7.13) obtain a parameter estimate for ϕ_3 that is just slightly more than twice its standard error. However, they do not employ a data transformation to remove heteroscedasticity and nonnormality and only use the Wolfer sunspot series from 1770 to 1869.

When one examines the ACF of the residuals of the AR(3) model fitted to the transformed series, one finds a large value at lag 9. This fact implies that it may be advisable to estimate a constrained AR(9) model without the parameters ϕ_3 to ϕ_8 included in the model. In Section 6.4.3, the Akaike information criterion (Akaike, 1974) also selects the constrained AR(9) model fitted to the sunspot series transformed using square roots as the best overall type of ARMA model to use. Previously, Schaerf (1964) also suggested modelling the sunspot data using a constrained AR(9) model but without the square root transformation.

5.5 OTHER IDENTIFICATION METHODS

5.5.1 Introduction

As demonstrated in the previous section, the sample ACF, PACF, IACF and IPACF are quite useful for ascertaining which subset of ARMA or ARIMA models are more suitable for fitting to a given time series. When these identification methods are used in conjunction with the Akaike information criterion (Akaike, 1974) in the manner described in Section 6.3, usually it is quite straightforward to select the most appropriate model. In addition, as exemplified by the application of ARMA modelling to the yearly sunspot series, the results of diagnostic checks can also be useful for iteratively designing the best specific model. Nonetheless, in most practical applications it is usually not necessary to employ other kinds of identification techniques beyond those described in Section 5.3. However, other identification methods are available and some of these procedures are now briefly outlined. Additional identification approaches are also mentioned in Section 6.3 where the Akaike information criterion is described.

5.5.2 R and S Arrays

Gray et al. (1978) develop a useful representation of the dependence structure of an ARMA process by transforming the theoretical ACF into two functions which they refer to as *R and S arrays*. In addition, Woodward and Gray (1979) define improved versions of these arrays, called the shifted R and S arrays. Moreover, Woodward and Gray (1979) present the generalized partial autocorrelation function as a related approach for model identification.

The R and S arrays are used for determining the orders of the operators in an ARIMA(p,d,q) model. Although it is usually more informative to display the identification results in tabular form, the R and S arrays can also be presented graphically. Computer programs are listed in the paper of Gray et al. (1978) for calculating the R and S plots. Moreover, Gray et al. (1978) present numerous practical applications while Woodward and Gray (1978) identify an ARMA(8,1) model to fit to yearly sunspot numbers by using the R and S arrays. The R and S arrays could be extended for identifying the seasonal ARIMA models of Chapter 12.

Salas and Obeysekera (1982) demonstrate the use of the generalized partial autocorrelation function as well as the R and S arrays for identifying ARMA models to fit to hydrologic time series. Furthermore, they present some recursive relationships for calculating the aforesaid identification methods.

5.5.3 The Corner Method

Beguin et al. (1980) present theoretical results for an identification procedure to ascertain the orders of p and q in an ARMA(p,q) model. To determine the orders of the AR and MA operators, the entries of what is termed a "Δ-array" are examined. Depending upon the form of the model, zero entries occur in a corner of the Δ-array according to a specified pattern and hence the approach is called the *corner method*. Beguin et al. (1980) claim that their identification methods are much simpler to use than those proposed by Gray et al. (1978).

5.5.4 Extended Sample Autocorrelation Function

Tsay and Tiao (1984) develop a unified approach for specifying the order of the operators required in an ARIMA(p,d,q) model to fit to a given time series. First, they propose an iterative regression procedure for obtaining consistent least square estimates for the AR parameters. Next, based upon the consistent AR estimates produced by iterated regressions, they define an *extended sample autocorrelation function* for use in model identification. The extended sample autocorelation function can be used to decide upon the order of differencing and also the numbers of AR and MA parameters that are needed. In practical applications, the authors propose that the calculations for the extended sample autocorrelation function be displayed in tabular form for conveniently identifying the ARIMA model.

5.6 CONCLUSIONS

The stages for constructing a time series model to fit to a given data set are portrayed in Figure III.1. At the identification stage, one must decide upon the parameters required in a model for fitting to a given data set. In particular, when designing an ARMA or ARIMA model to describe a nonseasonal time series, one must select the order of differencing as well as the number of AR and MA parameters that are needed. After examining a plot of the data, the parameters needed in the model can be ascertained by examining graphs of the sample ACF, PACF, IACF and IPACF, presented in Sections 5.3.4 to 5.3.7, respectively. Applications for illustrating how model identification is carried out in practice are presented in Section 5.4 for the cases of an annual riverflow series and a yearly sunspot series.

After identifying one or more tentative models to fit to the time series under consideration, one can obtain efficient estimates for the model parameters. In Chapter 6, the method of maximum likelihood is recommended for estimating the AR and MA parameters after any nonstationarity has been removed by differencing. Additionally, it is explained in Chapter 6 how an information criterion can be used to select the best model subsequent to estimating the parameters for more than one model. Finally, in Chapter 7 diagnostic checks are presented for deciding upon whether the fitted model adequately describes the time series. As shown in Figure III.1, when the model possesses inadequacies one can return to the identification stage in order to design an improved model which overcomes any difficulties. Usually, the results from the diagnostic check stage can be used for designing this improved model.

PROBLEMS

5.1 Three types of hydrological uncertainties are mentioned in Section 5.2.2. By referring to the references given in that section, explain in your own words what these uncertainties mean to you. Enhance your presentation by referring to specific examples of these uncertainties.

5.2 Summarize the types of modelling errors discussed by Warwick (1989). Compare these errors to the kinds of uncertainties discussed in Section 5.2.2.

5.3 Outline the unified description of model discrimination developed by Ljung (1978) and referenced in Section 5.2.3.

5.4 Suggest other types of modelling principles that are not mentioned in Section 5.2.4.

5.5 In Section 5.3.3 a list is presented for the kinds of information that may be found from examining a graph of a given time series. Describe three other benefits that may be realized by plotting observations over time.

5.6 Examine the graph of a time series which is of direct interest to you. Describe general statistical properties of the series that you can detect in the graph.

5.7 In Section 5.3.6, the IACF is defined. Why does Cleveland (1972) recommend using the sample ACF and IACF for ARMA model identification rather than the sample ACF and PACF? Summarize Chatfield's (1979) viewpoint about the use of the IACF in practical applications.

5.8 The original definition of the IPACF was made by Hipel et al. (1977) and is given in Section 5.3.7. Summarize Bhansali's (1983c) contributions to the development of the IPACF as an identification tool.

5.9 Develop the equations for determining the theoretical ACF and PACF for an ARMA(1,1) process.

5.10 Write down the equations for the theoretical IACF and IPACF for an ARMA(1,1) process.

5.11 Select a nonseasonal time series that is of interest to you. Obtain each of the identification graphs in Sections 5.3.3 to 5.3.7 for the series. Based upon these identification results, what is the most appropriate type of ARMA or ARIMA model to fit to the data set?

5.12 Using equations in your explanation, outline the approach of Gray et al. (1978) for determining the orders of the operators in an ARIMA(p,d,q) model. List advantages and drawbacks to their identification procedure.

5.13 Employing equations where necessary, describe the extended sample autocorrelation technique of Tsay and Tiao (1984) for identifying an ARIMA model. Discuss the advantages and disadvantages of their method as compared to its competitors.

REFERENCES

DATA SETS

Beveridge, W. H. (1921). Weather and harvest cycles. *Economics Journal*, 31:429-552.

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society*, 79:242-261.

Salas, J. D. and Yevjevich, V. (1972). Stochastic structure of water use time series. Hydrology Paper No. 52, Colorado State University, Fort Collins, Colorado.

Stokes, M. A., Drew, L. G. and Stockton, C. W. (1973). Tree ring chronologies of western America. Chronology Series 1, Laboratory of Tree Ring Research, University of Arizona, Tucson, Arizona.

United States Bureau of the Census (1976). *The Statistical History of the United States from Colonial Times to the Present*. Washington, D.C.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

DATA TRANSFORMATIONS

Fama, E. F. and Roll, R. (1968). Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, 63:817-837.

Fama, E. F. and Roll, R. (1971). Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66:331-339.

Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society*, Series B, 38(2):189-203.

Granger, C. W. J. and Orr, D. (1972). Infinite variance and research strategy in time series analysis. *Journal of the American Statistical Association*, 67(338):275-285.

Rosenfeld, G. (1976). Identification of time series with infinite variance. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 25(2):147-153.

Rothenburg, T. J., Fisher, F. M. and Tilanus, C. B. (1964). A note on estimation from a Cauchy distribution. *Journal of the American Statistical Association*, 59:460-463.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33:1-67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

MODEL IDENTIFICATION

Abraham, B. and Ledolter, J. (1984). A note on inverse autocorrelations. *Biometrika*, 71(3):609-614.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716-723.

Battaglia, F. (1988). On the estimation of the inverse correlation function. *Journal of Time Series Analysis*, 9:1-10.

Beguin, J. M., Gourieroux, C. and Monfort, A. (1980). Identification of a mixed autoregressive-moving average process: the Corner method. In Anderson, O. D., Editor, *Time Series*, pages 423-436, Amsterdam. North Holland.

Bhansali, R. J. (1980). Autoregressive and window estimators of the inverse correlation function. *Biometrika*, 67:551-566.

Bhansali, R. J. (1983a). A simulation study of autoregressive and window estimators of the inverse correlation function. *Applied Statistics*, 32(2):141-149.

Bhansali, R. J. (1983b). Estimation of the order of a moving average model from autoregressive and window estimates of the inverse correlation function. *Journal of Time Series Analysis*, 4:137-162.

Bhansali, R. J. (1983c). The inverse partial autocorrelation of a time series and its applications. *Journal of Multivariate Analysis*, 13:310-327.

Chatfield, C. (1979). Inverse autocorrelations. *Journal of the Royal Statistical Society*, Series A, 142:363-377.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics*, 14(2):277-298.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press, New York.

Gray, H. L., Kelley, G. D. and McIntire, D. D. (1978). A new approach to ARMA modelling. *Communications in Statistics*, B7(1):1-77.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 1, model construction. *Water Resources Research*, 13(3):567-575.

Hosking, J. R. M. (1980). The asymptotic distribution of the sample inverse autocorrelations of an autoregressive-moving average process. *Biometrika*, 67(1):223-226.

Lettenmaier, D. P., Hipel, K. W. and McLeod, A. I. (1978). Assessment of environmental impacts, Part Two: Data collection. *Environmental Management*, 2(6):537-554.

McClave, J. T. (1975). Subset autoregression. *Technometrics*, 17(2):213-220.

McLeod, A. I. (1984). Duality and other properties of multiplicative autoregressive-moving average models. *Biometrika*, 71:207-211.

McLeod, A. I., Hipel, K. W., and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, applications. *Water Resources Research*, 13(3):577-586.

Pagano, M. (1972). An algorithm for fitting autoregressive schemes. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 21:274-281.

Parzen, E. (1974). Some recent advances in time series modeling. *IEEE Transactions on Automatic Control*, AC-19(6):723-730.

Potter, K. W. (1976). Evidence for nonstationarity as a physical explanation of the Hurst phenomenon. *Water Resources Research*, 12(5):1047-1052.

Salas, J. D. and Obeysekera, J. T. B. (1982). ARMA model identification of hydrologic time series. *Water Resources Research*, 18(4):1011-1021.

Shaman, P. (1975). An approximate inverse for the covariance matrix of moving average and autoregressive processes. *Annals of Statistics*, 3:532-538.

Tsay, R. S. and Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association*, 79:84-96.

Warwick, J. J. (1989). Interplay between parameter uncertainty and model aggregation error. *Water Resources Bulletin*, 25(2):275-283.

Woodward, W. A. and Gray, H. L. (1978). New ARMA models for Wolfer's sunspot data. *Communication in Statistics*, B7(1):97-115.

Woodward, W. A. and Gray, H. L. (1979). On the relationship between the R and S arrays and the Box-Jenkins method of ARMA model identification. Technical Report Number 134, Dept. of Statistics, Southern Methodist University, Dallas, Texas.

MODELLING PHILOSOPHIES

Beck, M. B. (1987). Water quality modeling: A review of the analysis of uncertainty. *Water Resources Research*, 23(8):1393-1442.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Mass.

Caines, P. E. (1976). Prediction error identification methods for stationary stochastic processes. *IEEE Transactions on Automatic Control*, AC-21(4):500-506.

Caines, P. E. (1978). Stationary linear and nonlinear system identification and predictor set completeness. *IEEE Transactions on Automatic Control*, AC-23(4):583-594.

Hipel, K. W. (1993). Philosophy of modeling building. In Marco, J. B., Harboe, R. and Salas, J. D., editors, *Stochastic Hydrology and its Use in Water Resources Systems Simulation and Optimization*, Proceedings of the NATO (North Atlantic Treaty Organization) Advanced Study Institute on Stochastic Hydrology and its Use in Water Resources Simulation and Optimization, held Sept. 18-29, 1989, in Peniscola, Spain, 25-45, Dordrecht, the Netherlands. Kluwer Academic Publishers.

Jackson, B. B. (1975). The use of streamflow models in planning. *Water Resources Research*, 11(1):54-63.

Kashyap, R. L. and Rao, A. R. (1976). *Dynamic Stochastic Models from Empirical Data*. Academic Press, New York.

Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*. The Iowa State University Press, Ames, Iowa.

Kisiel, C. and Duckstein, L. (1972). Model choice and validation. In General Report, *Proceedings of the International Symposium on Uncertainties in Hydrologic and Water Resource Systems*, 1282-1308, Tucson, Arizona.

Ljung, L. (1978). Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, AC-23(5):770-783.

Tao, P. C. and Delleur, J. W. (1976). Multistation, multiyear synthesis of hydrologic time series by disaggregation. *Water Resources Research* 12(6):1303-1312.

Vicens, G. J., Rodriguez-Iturbe, I. and Schaake Jr., J. C. (1975). Bayesian generation of synthetic streamflows. *Water Resources Research*, 11(6):827-838.

Wood, E. F. (1978). Analyzing hydrologic uncertainty and its impact upon decision making in water resources. *Advances in Water Resources*, 1(5):299-305.

Wood, E. F. and Rodriguez-Iturbe, I. (1975). Bayesian inference and decision making for extreme hydrologic events. *Water Resources Research*, 11(4):533-542.

SUNSPOT NUMBER MODELS

Bailey, M. J. (1965). Prediction of an autoregressive variable subject both to disturbances and to errors of observation. *Journal of the American Statistical Association*, 60:164-181.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Craddock, J. M. (1967). An experiment in the analysis and prediction of time series. *The Statistician*, 17:257-268.

Davis, W. M. (1979). Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics*, 21(3):313-320.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysics Journal*, 126:152-158.

Moran, P. A. P. (1954). Some experiments in the prediction of sunspot numbers. *Journal of the Royal Statistical Society*, Series B, 16(1):112-117.

Morris, J. (1977). Forecasting the sunspot cycle. *Journal of the Royal Statistical Society*, Series A, 140(4):437-468.

Phadke, M. S. and Wu, S. M. (1974). Modelling of continuous stochastic processes from discrete observations with applications to sunspots data. *Journal of the American Statistical Association*, 69:325-329.

Schaerf, M. C. (1964). Estimation of the covariance and autoregressive structure of a stationary time series. Technical report, Department of Statistics, Stanford University, Stanford, California.

Schneider, S. H. and Mass, C. (1975). Volcanic dust, sunspots and temperature trends. *Science*, 190(4216):741-746.

Whittle, P. (1954). A statistical investigation of sunspot observations with special reference to H. Alfren's sunspot model. *Astrophysics Journal*, 120:251-260.

Woodward, W. A. and Gray, H. L. (1978). New ARMA models for Wolfer's sunspot data. *Communication in Statistics*, B7(1):97-115.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer sunspot numbers. *Phil. Transactions of the Royal Society*, Series A, 226:267-298.

CHAPTER 6

PARAMETER ESTIMATION

6.1 INTRODUCTION

In Chapter 5, a range of informative graphical methods are presented for identifying the parameters to include in an ARMA model for fitting to a given nonseasonal time series. As explained in that chapter, prior to deciding upon the form of the ARMA model, the data may be transformed using the Box-Cox transformation in [3.4.30] in order to alleviate problems with nonnormality and/or changing variance. Additionally, the differencing given in [4.3.3] may be required for removing nonstationarity. Whatever the case, at the identification stage, one must select one or more appropriate ARMA models from [3.4.4] or [4.3.4] for fitting to either the series as given or some modified version thereof.

As shown in Figure III.1, subsequent to identifying one or more tentative models for fitting to a particular series, one must obtain estimates for the parameters in the models. The main objectives of Chapter 6 are to present procedures for *estimating the parameters in ARMA models* and to explain how an automatic selection criterion such as the *Akaike information criterion* (Akaike, 1974) can be employed for choosing the best overall model when more than one model is calibrated.

For an identified ARMA model in [3.4.4] or [4.3.4], the following parameters must be estimated using the available data:

1. mean of the series,
2. AR parameters,
3. MA parameters,
4. innovation series,
5. variance of the innovations.

Because one often knows a priori the best type of *Box-Cox transformation* to use with a given kind of time series such as annual riverflows, one can first fix λ in [3.4.30] at a specified value before estimating the model parameters mentioned above. If λ is not known, it is possible to estimate λ along with the other model parameters. However, this requires a significant increase in the amount of computer time needed to estimate all the model parameters. Finally, one should keep in mind that the integer value for the *differencing parameter d* contained in ARIMA(p,d,q) models in Chapter 4 is selected using identification methods (see Sections 5.3.3 and 5.3.4). If differencing is used, often one may wish to fix the mean of the differenced series at zero and not estimate it (see discussion in Section 4.3.1). When d is allowed to take on real values to form the fractional ARMA models described in Chapter 11, one must estimate the value of d.

A given time series is just one possible realization or set of measurements from the phenomenon that generated it (see discussion in Sections 2.2 and 2.3). Because a time series contains only partial information about the phenomenon under study, the true or population values of the parameters of a model fitted to the series are not known. Consequently, there is *uncertainty* about the estimation of the model parameters. As explained in Section 6.2.3 and

Appendix A6.2, the uncertainty for a specified parameter estimate is quantified by what is called the *standard error* (SE) of the estimate.

Estimation theory was initiated by the great German mathematician Karl Friederich Gauss who developed the method of least squares for solving practical problems. Since the time of Gauss, well known researchers such as Sir R.A. Fisher, Norbert Wiener and R.E. Kalman, have developed an impressive array of estimation procedures and associated algorithms. These general approaches from estimation theory have been formulated for use with specific families of statistical models. For example, in this chapter the method of *maximum likelihood* is described and used for estimating the parameters of ARMA models. In Section 3.2.2, the Yule-Walker equations given in [3.2.11] can be employed for obtaining what are called *moment estimates* for AR models.

A great number of textbooks and research papers about estimation theory are available. Mendel's (1987) book, for example, covers a wide variety of estimation techniques including least squares, maximum likelihood and the Kalman filter (Kalman, 1960) approaches. A research paper by Norden (1972, 1973) presents a survey of maximum likelihood estimation which was originally developed by Fisher (1922, 1925). The monograph of Edwards (1972) also deals with the maximum likelihood approach to estimation. Most textbooks, in statistics, such as the ones by Kempthorne and Folks (1971) and Cox and Hinkley (1974) contain large sections dealing with estimation. In addition, statistical encyclopediae (Kotz and Johnson, 1988; Kruskal and Tanur, 1978; Kendall and Buckland, 1971) and handbooks (Sachs, 1984) have good explanations about estimation procedures.

Because of many attractive theoretical properties, maximum likelihood estimation is the most popular general approach to parameter estimation. In the next section, some of these properties are pointed out and maximum likelihood estimation for calibrating ARMA models is described. Subsequent to this, it is explained how the Akaike information criterion (Akaike, 1974) can be used to select the overall best model when more than one model is fitted to a specified time series. Practical applications are used for illustrating how estimation is carried out in practice and the Akaike information criterion can be used for model discrimination.

6.2 MAXIMUM LIKELIHOOD ESTIMATION

6.2.1 Introduction

The *probability distribution function* (pdf) of a set of random variables is written as a function of these variables and certain given parameters. For example, for the case of a single random variable following a normal distribution, the pdf is a function of this random variable and the parameters in the pdf are the mean and variance. When the actual values of the mean and variance are known for the normally distributed random variable, one can calculate the probability that the random variable takes on a value within a specified range by integrating the pdf over this range. On the other hand, if the measurements for a random variable are substituted into the pdf and the pdf is then considered as a function of the parameters that have not been estimated, the likelihood function is created. In other words, the *likelihood function* is essentially the probability of the actual data as a function of the parameters.

To be more specific, suppose that one is dealing with the sequence of observations in [4.3.3] which consists of n values represented by the vector $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$. The sample of n observations, $\mathbf{w}$, can be associated with an n-dimensional random variable having a known pdf, $p(\mathbf{w}|\beta)$, which depends on a vector of unknown parameters β. For the case of the ARMA model in [3.4.4] or the ARIMA model in [4.3.4], the parameters contained in β are the p AR parameters $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$, q MA parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, and the variance, σ_a^2, of the innovations. Hence, $\beta = (\phi, \theta, \sigma_a^2)$.

In advance of having the data, the pdf given by $p(\mathbf{w}|\beta)$ associates a density with a possible realization of $\mathbf{w}$, for fixed β. When the observations are available, one would like to find out values of β which could have produced the set of time series entries, $\mathbf{w}$. To accomplish this, one substitutes the data, $\mathbf{w}$, into the pdf and considers β as the variable in order to produce the likelihood function $L(\beta|\mathbf{w})$. Because of the way it is defined, the likelihood function has the same form as $p(\mathbf{w}|\beta)$. However, in the likelihood function the set of observations, $\mathbf{w}$, is fixed and the parameters contained in β are variable.

Because the relative value of the likelihood function, $L(\beta|\mathbf{w})$, is of main interest, the likelihood function often contains an arbitrary multiplicative constant. For simplifying calculations, it is usually more convenient to use the natural logarithm of the likelihood function given by $\ln L(\beta|\mathbf{w}) = l(\beta|\mathbf{w})$, which possesses an arbitrary additive constant. This function is commonly referred to as the *log likelihood function*.

In maximum likelihood estimation, one wants to determine the values of the parameters contained in β that maximize the likelihood function or, equivalently, the log likelihood function. These estimates are called *maximum likelihood estimates (MLE's)*.

One approach to finding the maximum value for a given function is to set the first order partial derivative with respect to each variable parameter equal to zero and then to solve these equations to find the values of the variables which maximize the function. Because the likelihood function for ARMA models is quite complicated, this simple approach cannot be used. Consequently, after defining the likelihood function for ARMA models in Section 6.2.3 and Appendix A6.1, some useful optimization algorithms are recommended for optimizing the likelihood function.

The second order partial derivative of the likelihood or log likelihood function with respect to each of the variable parameters reflects the rate of change of the slope or, in other words, the "spread" of the function. Accordingly, these second order derivatives, which are contained in the *information matrix* defined in Appendix A6.2, are used to determine approximate *standard errors (SE's)* for the MLE's. However, before going into the details of the likelihood function and the associated information matrix, the motivations for using maximum likelihood estimation are explained next.

6.2.2 Properties of Maximum Likelihood Estimators

Likelihood Principle

Prior to describing some of the attractive properties of maximum likelihood estimation, consider first an important characteristic of the likelihood function. One main reason why the likelihood function is of such great import in estimation theory is because of what is called the likelihood principle summarized below.

Likelihood Principle: Assuming that the underlying model is correct, all the information that the data can provide about the model parameters is contained in the likelihood function. All other aspects of the data are irrelevant with respect to characterizing the model parameters (Fisher, 1956; Barnard, 1949; Birnhaum, 1962).

The likelihood principle is in consonance with the Bayesian approach to statistics. This is because the likelihood function is a component in the posterior distribution of the parameters coming from the data.

As noted in the previous section, when the likelihood function, or equivalently, the log likelihood function is maximized, one obtains MLE's for the model parameters. The general mathematical expression which defines how one obtains MLE's for any set of data for a given family of models, is called the *maximum likelihood estimator*. In Appendix A6.1, for example, a maximum likelihood estimator is presented for calculating MLE's for the parameters of an ARMA model fitted to a given time series.

In general, most maximum likelihood estimators possess some fundamental statistical properties which, in turn, have led to the widespread development, acceptance and application of these estimators. To characterize estimators, Fisher (1925) introduced the concepts of consistency and large-sample efficiency. Although these concepts are defined in terms of large samples, estimators having these characteristics are usually well suited for use in practical applications. Because maximum likelihood estimators usually satisfy these concepts, Fisher and many other statisticians have advocated their employment for application purposes. The maximum likelihood estimators referred to in Section 6.2.3 and the one described in Appendix A6.1 are consistent and efficient. These two concepts are now briefly summarized. For detailed mathematical definitions of the concepts, the reader can refer to the references cited in this section as well as statistical encyclopediae, handbooks and standard textbooks.

Consistency

A *consistent estimator* is one which converges in probability as the sample size increases to the true value of the parameter or parameters being estimated. More specifically, let $\hat{\eta}$ be the estimate of a model parameter η using a given estimator for a sample size of n. The estimate $\hat{\eta}$ of η, or equivalently its estimator, is consistent if

$$\lim_{n\to\infty} P[|\hat{\eta}-\eta| > \epsilon] = 0 \qquad [6.2.1]$$

where P stands for probability, and ϵ is any positive number which can, of course, be very close to zero.

In practice, one would like to have an estimator which produces estimates that converge to the true values of the model parameters as the sample size increases. Although exceptions can be found, most maximum likelihood estimators are consistent.

Efficiency

Suppose that two estimators are consistent. Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the two consistent estimators or estimates for a model parameter η where the sample size is n. The *asymptotic relative efficiency (ARE)* of $\hat{\eta}_1$ with respect to $\hat{\eta}_2$ is:

$$ARE = \lim_{n \to \infty} \frac{var\ \hat{\eta}_1}{var\ \hat{\eta}_2} \qquad [6.2.2]$$

If the above ratio is less (more) than one, the estimator $\hat{\eta}_1$ is asymptotically more (less) efficient than $\hat{\eta}_2$ for estimating η. When the limit is equal to one, the estimators are equally efficient. The asymptotic relative efficiency is the limiting value of the *relative efficiency (RE)* given by:

$$RE = \frac{var\hat{\eta}_1}{var\hat{\eta}_2} \qquad [6.2.3]$$

For maximum likelihood estimators or MLE's, the variance of the MLE of a model parameter possesses minimum asymptotic variance and is asymptotically normally distributed when consistency and other conditions are satisfied (Cramer, 1946; Rao, 1973). Therefore, when investigating the properties of a given estimator, it is informative to compare it to its MLE counterpart. Suppose that $\hat{\eta}_1$ is the maximum likelihood estimator for a model parameter η and $\hat{\eta}_2$ is another estimator in [6.2.2]. Because the maximum likelihood estimator possesses minimum variance $0 \le ARE \le 1$. Furthermore, the ratio is referred to as the *first order asymptotic efficiency* of $\hat{\eta}_2$ with respect to the maximum likelihood estimator $\hat{\eta}_1$. If the first order efficiency is less than unity, the estimator $\hat{\eta}_2$ is less efficient than the maximum likelihood estimator $\hat{\eta}_1$ for large samples. However, when the first order efficiency is equal to one, the ratio in [6.2.2] cannot distinguish between the two estimators. One must then examine what is called second order efficiency (Rao, 1961, 1962) in order to select the most efficient estimator. *Second order efficiency* is concerned with the speed of convergence of the ratio in [6.2.2] and usually requires rather complicated expressions in order to be properly defined. Whatever the case, the maximum likelihood estimator is the only known estimator that possesses second order efficiency.

Gaussian Efficiency: In the definitions of the ARMA family of models in [3.4.4] and the ARIMA class of models in [4.3.4], the innovation series represented by a_t is assumed to be identically and independently distributed with a mean of zero and variance of σ_a^2 [i.e. IID(0,σ_a^2)]. To allow one to derive the likelihood function for these models, one must specify a distribution for the innovations. In practice, the a_t's are assumed to be normally independently distributed with a mean of zero and variance of σ_a^2 [i.e. NID(0,σ_a^2)]. The likelihood function for the Gaussian or normal case is discussed in Section 6.2.3 and presented in detail in Appendix A6.1. By determining the values of the model parameters which maximize the value of the likelihood or log likelihood function, one determines MLE's for the parameters. As explained in Appendix A6.2, the *covariance matrix* is obtained as the inverse of the *information matrix* and the entries along the diagonal give the variance of the estimates for the corresponding model parameters. The square root of these variances are called the SE's of estimation for the model parameters. Because the maximum likelihood procedure is used to obtain the parameter estimates, these SE's or, equivalently, the variances, possess Fisherian efficiency and, therefore, are the smallest values that can be obtained in large samples.

If the innovations are not normally distributed, one can use the same technique as for the NID case to obtain estimates for the model parameters. Even though the innovations do not follow a normal distribution, these estimates are called *Gaussian estimates* because the maximum likelihood estimator for NID innovations is used to calculate the estimates. It can be shown theoretically that the large sample covariance matrix for Gaussian estimates is the same as that for the situations for which the innovations are NID. This robustness property of maximum likelihood estimation under the normality assumption is referred to as *Gaussian efficiency* (Whittle, 1961; Hannan, 1970, pp. 377-383). However, the reader should keep in mind that even though the Gaussian estimates possess Gaussian efficiency, they are not Fisherian efficient (i.e. have minimum variances for the estimates) because the innovations do not follow a normal distribution.

Li and McLeod (1988) show how maximum likelihood may be used to fit ARMA models when the innovations, a_t, are non-Gaussian. For example, when the a_t are log-normal or gamma distributed, improved estimates of the parameters can be obtained by using maximum likelihood estimation.

6.2.3 Maximum Likelihood Estimators

A given nonseasonal series, z_t, may first be transformed using the Box-Cox transformation in [3.4.30] in order to make the series approximately normally distributed. Subsequent to a power transformation, the series can be differenced as in [4.3.3] just enough times to remove any nonstationarity. One then ends up with a stationary series w_t, $t = 1,2,\ldots,n$, which follows a normal distribution. By employing the identification procedures of Section 5.3, one can decide upon an appropriate ARMA(p,q) model to fit to the w_t series. Of course, if no Box-Cox transformation or differencing are needed, the w_t series is simply the original z_t observations.

Assuming that the innovations in [4.3.4] or [3.4.4] are NID, which also implies that the w_t or z_t sequences follow a normal distribution, one can derive the likelihood function for an ARMA model. By employing a suitable optimization algorithm to maximize the likelihood or log likelihood function with respect to the ARMA model parameters, one should theoretically be able to obtain MLE's for the parameters. However, the likelihood function is a fairly complicated expression and flexible algorithms are needed in order to make it computationally possible within a reasonable amount of time to maximize the likelihood function in order to find the MLE's. As a result, researchers have suggested saving computational time by maximizing approximations to the likelihood function to calculate *approximate MLE's* for the model parameters. As the sample size increases, the approximate MLE's approach closer and closer to the true MLE's. Box and Jenkins (1976, Ch. 7), for example, have put forward two approximate maximum likelihood procedures for ARMA models which are called the conditional and the unconditional or iterated methods. Generally speaking, their approaches do not work as well for ARMA models containing MA parameters and for time series that are fairly short (McLeod, 1977).

McLeod (1977) derives an approximate maximum likelihood procedure which is almost exact. His technique is referred to as the *modified sum of squares algorithm*. Besides providing parameter estimates that are very close to the true or exact MLE's, the approach is very efficient computationally and, therefore, requires relatively little computer time. Moreover, it works well with models containing MA parameters and series having relatively few observations.

More recently, a number of authors have developed *exact maximum likelihood estimators* for use with ARMA models. These exact techniques include contributions by:

1. Newbold (1974),
2. Ansley (1979),
3. Ljung and Box (1979), and
4. Mélard (1984) who uses a Kalman filter approach to maximum likelihood estimation.

As just noted, McLeod's (1977) estimation technique for ARMA models is computationally efficient and produces estimates that are almost exact MLE's. Furthermore, the procedure has been extended for use with seasonal ARMA models (McLeod and Sales, 1983). Accordingly, this flexible algorithm is recommended for use in practical ARMA modelling and is outlined in Appendix A6.1. The McLeod-Hipel time series package referred to in Section 1.7 contains the estimation algorithm of Appendix A6.1 as well as other approximate and exact maximum likelihood estimators.

In addition to possessing desirable statistical properties, maximum likelihood estimation is computationally convenient. This is because a range of useful and powerful optimization techniques are available to maximize or minimize a function such as the likelihood or log likelihood function with respect to the model parameters. Some of the optimization algorithms that have been extensively utilized in practical applications include:

1. Gauss linearization (Draper and Smith, 1980),
2. steepest descent (Draper and Smith, 1980),
3. Marquardt algorithm which is a combination of the above two algorithms (Marquardt, 1963),
4. conjugate directions (Powell, 1964, 1965).
5. Davidon's Algorithm (Davidon, 1968) for which a FORTRAN subroutine is provided by Ishiguro and Akaike (1989) for log likelihood maximization.

For the applications given in Section 6.4 and other chapters in this textbook conjugate directions is used in conjunction with the estimation procedure of Appendix A6.1 to obtain MLE's for the ARMA model parameters. For an explanation of the variety of optimization methods, the reader can refer to textbooks such as those by Luenberger (1984), Gill et al. (1981) and VanderPlaats (1984).

To obtain estimates for the parameters in an ARMA model, a time series of observations is used with an appropriate maximum likelihood estimator. Because this time series is only a finite sample realization of the phenomenon generating the series, the MLE for a given parameter is not the population value. The SE or standard derivation of the estimate is used to reflect the *uncertainty* contained in the estimate. In Appendix A6.2, it is explained how the SE's for the parameter estimates are defined. More specifically, the variance-covariance matrix of the parameter estimates is the inverse of what is called the information matrix. The square roots of the diagonal entries in the variance-covariance matrix provide the estimates of the SE's for the estimated model parameters. Furthermore, because it is known that MLE's are asymptotically normally distributed, one can obtain 95% confidence limits for a given parameter estimate. If for example, zero were contained within the interval formed by a parameter estimate ± 1.96SE, one could argue that the parameter estimate is not significantly different from zero and perhaps the parameter should be left out of the model.

6.3 MODEL DISCRIMINATION USING THE AKAIKE INFORMATION CRITERION

6.3.1 Introduction

As noted in Chapter 5, the practitioner is usually confronted with the problem of choosing the most appropriate model for fitting to a given data set from a large number of available models. Consequently, *model discrimination* procedures are required and some possible selection methods are listed in Section 5.2.3. The identification methods in Section 5.3 constitute graphical and tabular techniques that can assist in deciding upon which model to choose. However, these methods require some skill when being used in applications since the modeller must be cognizant of the properties of the various types of identification graphs in order to ascertain which parameters should be included in the model. To increase the speed, flexibility, accuracy and simplicity involved in choosing a model, the *Akaike Information Criterion (AIC)* (Akaike, 1974) has been found to be quite useful. The AIC was first suggested for use in hydrology by Hipel et al. (1977) and McLeod et al. (1977). Hipel (1981) explains in detail how the AIC can be used in geophysical model discrimination and provides references for its application to many different kinds of time series.

6.3.2 Definition of the Akaike Information Criterion

Based upon information theory, Akaike (1972a, 1973, 1974) developed the AIC which is defined as

$$AIC = -2\ln ML + 2k \qquad [6.3.1]$$

where ML denotes maximum likelihood, $\ln ML$ is the value of the maximized log likelihood function for a model fitted to a given data set, and k is the number of independently adjusted parameters within the model. A desirable attribute of the AIC is that the modelling principles described in Sections 1.3 and 5.2.4 are formally incorporated into the equation. The first term on the right hand side of [6.3.1] reflects the doctrine of *good statistical fit* while the second entry accounts for *model parsimony*. Because of the form of [6.3.1], when there are several available models for modelling a given time series, the model that possesses the minimum value of the AIC should be selected. This procedure is referred to by Akaike (1974) as *MAICE (minimum AIC estimation)*.

The original mathematical development for the AIC formula in [6.3.1] is given by Akaike (1973, 1974) while a summary of the derivation is presented by Ozaki (1977) and also Kitagawa (1979). Even though the entries in [6.3.1] reflect sound modelling principles, as noted by Akaike (1978) "... the only justification of its use will come from its performance in applications." The MAICE procedure has previously been successfully applied to a wide range of statistical problems. The method has been used to decide upon the order of an ARMA model to fit to a time series (Akaike, 1974; Hipel et al., 1977; McLeod et al., 1977; Ozaki, 1977), to ascertain the type of nonstationary ARIMA model to describe a time series (Ozaki, 1977), to determine the order of an AR model (Akaike, 1978; Akaike, 1979; Shibata, 1976), to select the order of a

Markov chain (Tong, 1975), to decide upon the order of a polynomial regression (Akaike, 1972b; Tanabe, 1974), to determine the number of factors needed in a factor analysis (Akaike, 1971), to assist in robot data screening (Akaike, 1972b), to detect outliers in a data set (Kitagawa, 1979), to analyze cross classified data (Sakamoto and Akaike, 1977), and to assist in canonical correlation analysis of time series (Akaike, 1976). The AIC can be employed to select the most suitable model when more than one family of models are being considered and McLeod and Hipel (1978) used the MAICE procedure to determine whether an ARMA or Fractional Gaussian noise model should be utilized to model a given annual hydrological time series (see Section 10.4). The AIC can be employed to select the best model from the families of seasonal models discussed in Part VI, and to choose the most appropriate intervention model (see Chapter 19). In fact, the MAICE procedure can be used with all the models considered in this book (see Table 1.6.2) and the wide range of applications presented by Hipel (1981) confirm the versatility of this method for selecting the most appropriate model to fit to a time series.

6.3.3 The Akaike Information Criterion in Model Construction

Employment of the MAICE procedure reinforces and complements the identification, estimation and diagnostic stages of model constructions illustrated in Figure III.1 at the start of Part III. Figure 6.3.1 depicts how MAICE can be incorporated into the three stages of model development. Even though this chapter is concerned with nonseasonal ARMA and ARIMA models, the same general methodology can be employed no matter what types of time series models are being considered. For instance, the AIC model building procedure is recommended for use with the long memory, seasonal, transfer function-noise, intervention and multivariate models of Parts V to IX, respectively.

As shown by the flow chart in Figure 6.3.1, there are basically two approaches for employing the MAICE procedure in model construction. One method is to calculate the AIC for all possible models which are considered worthwhile for fitting to a given data set. For example, after specifying the Box-Cox parameter λ in [3.4.30] (often λ is set equal to unity if it is not known beforehand that a transformation is needed) maximum values for p, q and perhaps d may be set for ARIMA(p,d,q) models. The AIC can then be calculated for all possible combinations of p, d and q and the ARIMA model with the minimum AIC value is chosen. Although the selected model can usually be shown to adequately satisfy the important residual assumptions, as shown in Figure 6.3.1 it is always advisable to check for whiteness, normality, and homoscedasticity of the residuals using the methods in Sections 7.3 to 7.5, respectively. When the residuals are not white, other models should be considered by specifying a more flexible range for p, d and q. If the residuals do not possess constant variance and perhaps are not normally distributed then a suitable Box-Cox transformation may rectify the situation. To select the most suitable value of λ, a range of values of λ may be tried for the best ARIMA(p,d,q) model which was just chosen using the MAICE procedure. The value of λ which minimizes the AIC for the ARIMA model is then chosen. Another method is to obtain a MLE of λ for the best model and then to fix λ at this value if the exhaustive enumeration is repeated. Of course, λ could be estimated for all possible combinations of p, d and q in the exhaustive enumeration, but this would require a very large amount of computer usage.

If the diagnostic check stage is skipped and information from the identification and estimation stages is ignored when employing the exhaustive enumeration procedure with the AIC, it is possible that the best model may be missed. For example, in Section 6.4.3 it is shown that the

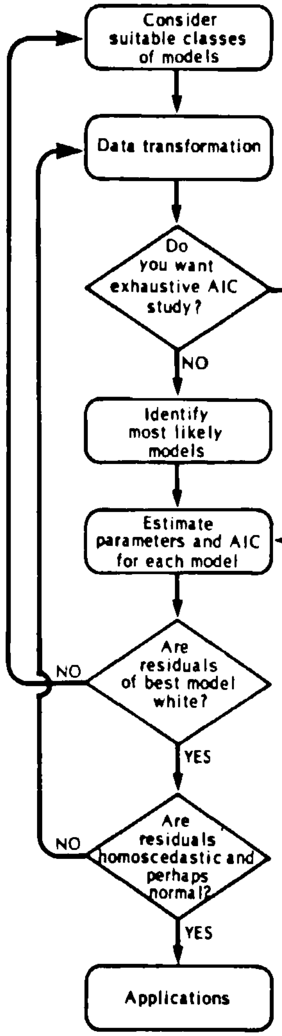

Consider suitable classes of models
Data transformation
Do you want exhaustive AIC study?
YES
NO
Identify most likely models
Specify range of parameters
Estimate parameters and AIC for each model
Are residuals of best model white?
NO
YES
Are residuals homoscedastic and perhaps normal?
NO
YES
Applications

Figure 6.3.1. Model construction using MAICE.

most suitable type of ARMA model to fit to the annual sunspot series is an AR model of order 9 with the third to eighth AR parameter omitted from the model and the data transformed by a square root transformation. As explained in that section, if diagnostic testing had not been done and the SE's of the parameter estimates had not been considered, the most suitable model would not have been discovered. Besides the annual sunspot series, the MAICE procedure is used in Section 6.4.2 to decide upon the most appropriate ARMA model to fit to the average annual flows of the St Lawrence River at Ogdensburg, New York.

An exhaustive AIC study may prove to be rather expensive due to the amount of computations. Consequently, as illustrated in Figure 6.3.1 an alternative approach is to only estimate the parameters and hence the AIC for a subset of models. For example, information from the identification stage (see Chapter 5) may indicate three tentative models to fit to the time series. The AIC is then only calculated for these three models and the model with the minimum AIC value is selected. If there are any problems with the residuals, appropriate action may be taken as shown in Figure 6.3.1. Otherwise, the chosen model can be employed in practical applications such as forecasting (see Chapter 8) or simulation (Chapter 9).

6.3.4 Plausibility

A question which is often asked by practitioners is how to interpret the relative differences in the values of the AIC for the various models which are fit to a specified data set. In fact, the different AIC values can be interpreted in a variety of manners. For example, if one model possesses an AIC value which is approximately $2k$ less than that of another model, this is analogous to the superior model having k less parameters than the other model. A lower AIC value can also be considered to be mainly due to a better statistical fit because of the first term on the right hand side of [6.3.1]. However, a lower AIC value is usually caused by both components of the formula in [6.3.1] and, therefore, an alternative approach for interpreting the differences in the AIC values between two models is to consider plausibility.

As shown by Akaike (1978), $\exp(-0.5AIC)$ is asymptotically a reasonable definition of the plausibility of a model specified by the parameters which are determined by the method of maximum likelihood. Consequently, the *plausibility* of model i versus model j can be calculated using

$$\text{Plausibility} = \exp[0.5(AIC_j - AIC_i)] \qquad [6.3.2]$$

where AIC_i is the value of the AIC for the ith model and AIC_j is the AIC value for the jth model. Table 6.3.1 displays some representative results for the plausibility of model i against model j where the jth model is assumed to have a lower AIC value than model i. As can be seen in Table 6.3.1, it is only the relative difference of the AIC values that is important and as these differences increase the plausibility decreases exponentially. Notice that when the AIC values differ by 6 the plausibility is only about 5%.

6.3.5 Akaike Information Criterion for ARMA and ARIMA Models

To determine the value of the AIC for an ARMA(p,q) model, both terms in [6.3.1] must be calculated separately. By optimizing the log likelihood function with respect to the model parameters (see Section 6.2.3 and Appendix A6.1), the value of the maximized log likelihood can be found for substitution into 6.3.1. The number of model parameters k is due to p AR parameters, q MA parameters, the variance of the model residuals, the Box-Cox exponent λ if it is included in the model, and the mean of the transformed series.

When considering a nonstationary series of length N, the data is differenced d times using [4.3.3] to produce a stationary series of length $n = N - d$. Because the differencing reduces the amount of information, this will certainly affect the first term on the right hand side of [6.3.1]. Hence, the AIC for an ARIMA model can be roughly calculated as

Table 6.3.1. Plausibility of model i versus model j.

$-(AIC_j - AIC_i)$	Plausibility
1	0.6065
2	0.3679
3	0.2313
4	0.1353
5	0.0821
6	0.0498
7	0.0302
8	0.0183
9	0.0111
10	0.0067
15	0.0006

$$AIC = \frac{N}{n}(-2\ln ML) + 2k \qquad [6.3.3]$$

where the value of the maximized log likelihood is obtained by optimizing the logarithm of [A6.1.5]. The total number of parameters k is the same as that for the ARMA model except when the mean of the differenced series is assumed to be zero and hence is not estimated, the number of parameters is decreased by one.

Another alternative for developing an AIC formula for an ARIMA model is to alter both components on the right hand side of [6.3.1]. As argued by Ozaki (1977), an increase in the number of data points contributes to decreasing the penalty due to the number of parameters. This effect can be incorporated into the AIC by writing the formula as

$$AIC = \frac{N}{n}(-2\ln ML + 2k) \qquad [6.3.4]$$

6.3.6 Other Automatic Selection Criteria

As shown by Figure 6.3.1 the MAICE procedure tends to "automate" model construction and to simplify model selection. In practice, it has been found that the MAICE methodology almost always chooses the same models which would be selected using more time consuming methods such as those presented in Section 5.3 and elsewhere. For example, when one model is a subset of another, a likelihood ratio test can be employed to determine if the model with more parameters is required for modelling a specified data set. However, as shown by McLeod et al. (1977), results from likelihood ratio tests usually confirm the conclusions reached using the MAICE procedure. An additional advantage of MAICE is that it is not necessary to select subjectively a significance level as is done with the likelihood ratio test.

The AIC is not the only *automatic selection criterion (ASC)* that can be used in model discrimination, although it is probably the most flexible and comprehensive of the methods which are presently available. For choosing the order of an AR model, Akaike (1969, 1970) developed the point estimation method called the final prediction error (FPE) technique (see Appendix A6.3 for a definition of the FPE and its relationship to the AIC). McLave (1975) presented an algorithm to be used in subset autoregression for obtaining the best constrained AR model where model selection is based upon the FPE criterion (see Section 3.4.4 for a discussion of constrained models). In another paper, McLave (1978) compared the FPE technique and a sequential testing approach which he referred to as the "max χ^2 method" for choosing the constrained AR model. Other ASC's which can only be used for AR modelling include the technique devised by Anderson (1971), the "CAT" criterion of Parzen (1974), and the method of Hannan and Quinn (1979). The "D-statistic" of Gray et al. (1978) can be utilized for choosing the most appropriate nonseasonal ARMA(p,q) model, although the statistic has not been sufficiently developed for use in nonstationary and seasonal modelling. Mallows (1973) developed a statistic for use in model discrimination that is related to what he calls the C_p statistic. Based upon the characteristics of the sample ACF and the sample PACF (refer to Sections 5.3.4 and 5.3.5 for explanations of the sample ACF and PACF, respectively), Hill and Woodworth (1980) employed a pattern recognition technique to identify the more promising ARIMA models that should be considered for fitting to a specified time series. Following this, they recommended using an appropriate ASC to select the overall best model. Akaike (1977), Rissanen (1978) and Schwarz (1978) developed similar selection criteria for use with ARMA models while Chow (1978) proposed an improved version of these methods. Sawa (1978) defined a criterion for statistical model discrimination called the minimum attainable Bayes risk. Stone (1979) compared the asymptotic properties of the AIC and Schwarz criterion while Hannan (1979, 1980) derived important theoretical results for various kinds of ASC's. Based on the Kullback Leibler information number, Shibata (1989) obtained the TIC (Takeuchi's Information Criterion) as a natural extension of the AIC. He then went on to develop the RIC (Regularization Information Criterion) as a meaningful expansion of both the AIC and TIC. Moreover, Shibata (1989) compared various ASC's in terms of criteria which include consistency and efficiency.

As pointed out in Section 1.3.3, many of the ASC's have a structure which is quite similar to that of the AIC in [6.3.1]. Consider, for instance, Schwarz's approximation of the *Bayes information criterion (BIC)* (Schwarz, 1978) which is written as

$$BIC = -2\ln ML + k\ln(n) \qquad [6.3.5]$$

As is also the case for the AIC in [6.3.1], the first term on the right hand side of [6.3.5] reflects good statistical fit while the second component is concerned with model parsimony. When fitting more than one model to a given time series, one selects the model which gives the lowest value of the BIC. To employ an ASC such as the BIC in [6.3.5] in model construction, simply replace the AIC by the other ASC in Figure 6.3.1. As explained in Section 6.3.3, there are two basic approaches for utilizing an ASC in model development.

Certainly further theoretical and practical research is required to compare the capabilities of the more promising automatic selection procedures. However, the efficacy of MAICE is clearly demonstrated by the many and varied applications cited in this book and elsewhere. For instance, MAICE can be employed to choose the best model from different families of seasonal models (see Part VI), and to design transfer-function noise (Part VII), intervention (Part VIII) and multivariate ARMA (Part IX) models. Furthermore, when considering different types of models for forecasting, usually the kind of model which forecasts most accurately also possesses the lowest AIC value (see Chapter 15). Some disadvantages of MAICE and the other ASC's are that an overall statistic tends to cover up much of the information in the data and the practitioner may lose his or her sense of feeling for the inherent characteristics of the time series if he or she bases his or her decisions solely upon one statistic. However, when MAICE is used in conjunction with the three stages of model construction as is shown in Figure 6.3.1, there is no doubt that MAICE greatly enhances the modelling process.

Akaike (1985) clearly explains how the derivation of the AIC is based upon the concept of *entropy*. In fact, the minimum AIC procedure can be considered as a realization of the *entropy-maximization principle* (Akaike, 1977). A further attractive theoretical feature of the MAICE approach is that it can be used to compare models which are not nested. Therefore, as noted earlier, one can use the MAICE procedure to select the best overall model across different families of models, as is done in Part VI for seasonal models. The practical import of the MAICE method for use in model discrimination is demonstrated by the two applications in the next section.

6.4 APPLICATIONS

6.4.1 Introduction

Table 5.4.1 in the previous chapter lists ARMA models identified for fitting to five nonseasonal stationary natural time series. In addition, Table 5.4.2 and Section 4.3.3 presents ARIMA models selected for fitting to three nonseasonal nonstationary time series. The maximum likelihood estimator described in Appendix A6.1 and mentioned in Section 6.2.3 can be used to calculate MLE's and SE's for the parameters in all of the foregoing models. Moreover, when more than one model is fitted to a given time series, the AIC of Section 6.3 can be employed for choosing the most appropriate model.

In the next two sections, estimation results along with applications of the AIC are presented for the same two case studies for which detailed identification findings are given in Section 5.4. The first application deals with modelling the average annual flows of the St. Lawrence River at Ogdensburg, New York, while the second one is concerned with modelling average annual sunspot numbers.

6.4.2 Yearly St. Lawrence Riverflows

Average annual flows for the St. Lawrence River at Ogdensburg, New York, are available from 1860 to 1957 (Yevjevich, 1963) and plotted in Figures 2.3.1 and 5.4.1 in m^3/s. The sample ACF, PACF, IACF, IPACF for these flows are displayed in Figures 5.4.2 to 5.4.5, respectively. As explained in Section 5.4.2, these identification graphs indicate that probably the best type of ARMA model to fit to the St. Lawrence flows is a constrained AR(3) model without the ϕ_2 parameter. However, one may also wish to try fitting AR(1) and unconstrained AR(3) models.

Table 6.4.1 lists the MLE's and SE's for AR(1), AR(3) and constrained AR(3) models fitted to the St. Lawrence flows. The theoretical definition for AR models can be found by referring to [3.2.5].

Model discrimination can be accomplished by comparing parameter estimates to their SE's, by using the AIC or by performing the likelihood ratio test. In order to employ the first procedure, first consider the models listed in Table 6.4.1. Notice that for both the AR(3) model and the AR(3) model without ϕ_2 the estimate $\hat{\phi}_3$ for ϕ_3 is more than twice its standard error. Therefore, it can be argued that even at the 1% significance level, ϕ_3 is significantly different from

Table 6.4.1. Parameter estimates for the AR models fitted to the annual St. Lawrence riverflows.

Models	Parameters	MLE's	SE's	AIC's
AR(1)	ϕ_1	0.708	0.072	1176.38
	σ_a	419.73		
AR(3)	ϕ_1	0.659	0.099	1175.59
	ϕ_2	-0.087	0.119	
	ϕ_3	0.216	0.099	
	σ_a	409.15		
Constrained AR(3) without ϕ_2	ϕ_1	0.619	0.084	1174.11
	ϕ_3	0.177	0.084	
	σ_a	410.27		

zero and should be included in the model. Consequently, the AR(1) model should not be utilized to model the St. Lawrence riverflows. Furthermore, because the SE for $\hat{\phi}_2$ in the AR(3) model is greater than $\hat{\phi}_2$, for model parsimony the AR(3) model without ϕ_2 is the proper model to select.

When the AIC is employed for model selection, it is not necessary to choose subjectively a significance level, as is done in hypothesis testing. By using [6.3.1], the values for the AIC are calculated for the three AR models and listed in the right hand column of Table 6.4.1. As can be seen, the AR(3) model without ϕ_2 has the minimum AIC, and, therefore, the AIC also indicates that this model should be chosen in preference to the others.

Suppose that one wishes to discriminate between models where one model is a subset of another. For the case of an AR model, let the order of one AR model be k and the order of another model containing more AR parameters be r. Let the residual variances of these two models be $\hat{\sigma}_a^2(k)$ and $\hat{\sigma}_a^2(r)$, respectively. The *likelihood ratio statistic* given by

$$n\ln\left[\hat{\sigma}_a^2(k)/\hat{\sigma}_a^2(r)\right] \sim \chi^2(r-k) \qquad [6.4.1]$$

is χ^2 distributed with $r-k$ degrees of freedom. If the calculated $\chi^2(r-k)$ from [6.4.1] is greater than $\chi^2(r-k)$ from the tables at a chosen significance level, a model with more parameters is needed.

The above likelihood ratio can be utilized to choose between the AR(1) model and the AR(3) model with $\phi_2=0$. By substituting $n=97$, $k=1$, the residual variance of the AR(1) model for $\hat{\sigma}_a^2(k)$, $r=2$, and the residual variance of the AR(3) model with $\phi_2=0$ for $\hat{\sigma}_a^2(r)$, the calculated χ^2 statistic from [6.4.1] has a magnitude of 4.58. For 1 degree of freedom, this value is significant at the 5% significance level. Therefore, this test indicates that the constrained AR(3) model should be selected in preference to the AR(1) model.

The likelihood ratio test can also be employed to test whether an AR(3) model without ϕ_2 gives as good a fit as the AR(3) model. Simply substitute into [6.4.1] $n=97$, $k=2$, the residual variance of the AR(3) model with $\phi_2=0$ for $\hat{\sigma}_a^2(k)$, $r=3$, and the residual variance of the AR(3) model for $\hat{\sigma}_a^2(r)$. The calculated χ^2 statistic possesses a value of 0.0569. For 1 degree of freedom this value is certainly not significant even at the 50% significance level. Consequently, the constrained model without ϕ_2 gives an adequate fit and should be used in preference to the AR(3) model in order to achieve model parsimony.

By substituting the estimated AR parameters into [3.2.5], one can write the constrained AR(3) model without ϕ_2 as

$$(1 - 0.619B - 0.177B^3)(z_t - 6819) = a_t \qquad [6.4.2]$$

where z_t is the average annual flow at time t, and 6819 is the MLE of the mean for the z_t series. Diagnostic checks presented in the next chapter in Section 7.6.2 demonstrate that the constrained AR(3) model without ϕ_2 adequately models the average annual flows of the St. Lawrence River.

6.4.3 Annual Sunspot Numbers

The yearly Wolfer sunspot number series is available from 1700 to 1960 (Waldmeier, 1961) where a plot of the series from 1770 to 1869 is shown in Figure 5.4.6. The sample ACF, PACF, IACF and IPACF are presented in Figures 5.4.7 to 5.4.10 in the identification chapter. As explained in Section 5.4.3, these identification graphs in conjunction with the output from diagnostic checks (see Section 7.6.3) indicate that an appropriate model may be a constrained AR(9) model without ϕ_3 to ϕ_8 fitted to the square roots of the sunspot series.

The MAICE procedure of Section 6.3 can be used to select the best type of ARMA model to fit to the sunspot series. Previously, Ozaki (1977) found using MAICE that an ARMA(6,3) model is the most appropriate model to fit to the given sunspot series having no data transformation. Akaike (1978) employed the AIC to select an ARMA(7,3) model with a square root transformation as the best sunspot model. However, Akaike (1978) did note that, because of the nature of sunspot activity, a model based on some physical consideration of the generating mechanism may produce a better fit to the data. Nevertheless, in this section it is shown how the model building procedure outlined in Figure 6.3.1 can be used to select an even better model from the family of ARMA models. As was suggested by McLeod et al. (1977) and also Hipel (1981), the AR(9) model, with a square-root transformation and the third to eighth AR parameters omitted from the model, produces a lower value of the AIC than all of the other aforementioned ARMA models. Earlier, Schaerf (1964) suggested modelling the sunspot data using a constrained AR(9) model but without the square-root transformation.

Because Ozaki (1977) used the series of 100 sunspot values listed as series E in the book of Box and Jenkins (1976), the same data set is used here for comparison purposes. By using an exhaustive enumeration procedure, Ozaki (1977) calculated the AIC for all ARMA(p,q) models for $0 \le p,q \le 9$ and found that an ARMA(6,3) model possessed the minimum AIC value. Employing [6.3.1], the values of the AIC were calculated for the same set of models examined by Ozaki (1977). The second column of Table 6.4.2 lists the AIC values for some of the models when the data is not transformed using [3.4.30] (i.e., $\lambda=1$ and $c=0$ in [3.4.30]). It can be seen that the minimum AIC value occurs for the ARMA(6,2) model, which is almost the same as the value of the AIC for the constrained AR(9) model. Notice that the ARMA(6,3) model suggested by Ozaki (1977) has a much higher AIC value than those for the ARMA(6,2) and constrained AR(9) models. This discrepancy is probably due to the different estimation procedure used by Ozaki.

The estimation method of McLeod (1977) described in Appendix A6.1 provides parameter estimates that are closer approximations than those of Box and Jenkins (1976) to the exact MLE's. As shown by McLeod (1977), implementation of his estimation method can result in improved estimates of the model parameters especially when MA parameters are contained in the model. As far as Table 6.4.2 is concerned, the improved estimation procedure affects the log likelihood in [6.3.1] and this in turn causes the AIC values to be slightly different than those given by Ozaki (1977, p. 297, Table 6).

Because information from the three stages of model construction is essentially ignored when using an exhaustive AIC enumeration such as the one adopted by Ozaki (1977), the best ARMA model is missed. To avoid this type of problem, the AIC can be combined with model construction as shown in Figure 6.3.1. From the plots of the sample ACF, PACF, IACF, and IPACF in Figures 5.4.7 to 5.4.10, respectively, it is difficult to decide upon which model to estimate. However, the sample PACF does possess values at lags 1 and 2 which are significantly different from zero and also some rather larger values at lags 6 to 9. When an ARMA(2,0) model is fitted to the data the independence, normality and homoscedastic assumptions (see Sections 7.3 to 7.5, respectively) are not satisfied. The residual ACF (see Section 7.3.2) has a large value at lag 9 and this fact suggests that an AR parameter at lag 9 should perhaps be incorporated into the model. The value of the AIC is lowest in column 2 of Table 6.4.2 for the AR(9) model without AR parameters from lags 3 to 8. However, because the statistic for changes in residual variance depending on the current level of the series, the statistic for trends in variance over time (see Section 7.5.2) and the skewness coefficient (see Section 7.4.2) all possess magnitudes which are more than twice their standard error, this points out the need for a Box-Cox transformation to eliminate heteroscedasticity and nonnormality. A square-root transformation can be invoked by setting λ equal to 0.5 in [3.4.30] and assigning the constant c a value of 1.0 due to the zero values in the sunspot series. Notice from the entries in the third column in Table 6.4.2 that a square-root transformation drastically lowers the AIC values for all of the models. The best model is an AR(9) or ARMA(9,0) model with a square-root transformation and without the third to eighth AR parameters. This constrained model was not missed because information from the model construction stages was used in conjunction with the MAICE procedure. Hence, when modelling a complex time series such as the sunspot data, it is advantageous for the practitioner to interact at all stages of model development by following the logic in Figure 6.3.1.

In Table 6.4.3, the MLE's and SE's are shown for the parameters of the most appropriate ARMA model which is fitted to the sunspot time series. When considering the 100 observations from 1770 to 1869 which are listed as Series E in Box and Jenkins (1976), the difference equation for the constrained AR(9) model with a square-root transformation is written as

$$(1 - 1.325B + 0.605B^2 - 0.130B^9)(w_t - 10.718) = a_t \qquad [6.4.3]$$

where

$$w_t = (1/0.5)[(z_t + 1.0)^{0.5} - 1.0]$$

is the transformation of the given z_t series for the sunspot numbers. The calibrated difference equation for the model fitted to the entire sunspot series from 1700 to 1960 is

Table 6.4.2. AIC values for the ARMA sunspot models.

ARMA (p,q) Model	AIC for λ=1, c=0.0	AIC for λ=0.5, c=1.0
(1,0)	618.30	580.40
(2,0)	551.85	518.41
(2,1)	547.63	519.19
(3,0)	549.57	519.13
(4,4)	546.98	516.10
(5,1)	547.29	523.82
(5,4)	548.20	517.96
(5,5)	547.23	517.38
(6,1)	548.11	517.39
(6,2)	545.05	518.43
(6,3)	551.35	523.34
(6,4)	550.73	502.40
(7,1)	548.17	518.98
(7,3)	551.01	521.37
(8,0)	545.67	519.83
(8,1)	547.65	520.44
(9,0)	547.65	519.72
(9,1)	548.18	518.78
Constrained (9,0)	545.64	511.58

$$(1 - 1.245B + 0.524B^2 - 0.192B^9)(w_t - 10.673) = a_t \qquad [6.4.4]$$

As shown in Section 7.6.3, the sunspot model in [6.4.4] satisfies diagnostic checks.

Table 6.4.3. Parameter estimates for the constrained AR(9) model fitted to the square roots of the yearly sunspot observations.

Parameters	MLE's	SE's
ϕ_1	1.325	0.074
ϕ_2	-0.605	0.076
ϕ_9	0.130	0.042
μ	10.718	1.417
σ_a^2	4.560	

By using [6.3.2] the relative plausibility of the sunspot models can be obtained. For instance, from Table 6.4.2 the next best model to the one in [6.4.3], according to the AIC, is an ARMA(4,4) model with a square-root transformation. When the appropriate AIC values from Table 6.4.2 are substituted into [6.3.2], the plausibility of the ARMA(4,4) model with a square-root transformation as compared to the best model is 0.10. According to the AIC, all of the other ARMA models with a square-root transformation are less plausible than even the ARMA(4,4) model. In addition, a comparison of the entries in columns two and three of Table 6.4.2 reveals how a square-root transformation significantly lowers the AIC values and hence increases the plausibility of a given ARMA model.

6.5 CONCLUSIONS

As explained in Section 6.2.2, maximum likelihood estimators possess a range of very desirable statistical properties which makes them highly attractive for use in practical applications. For example, maximum likelihood estimators are efficient and therefore produce parameter estimates having minimum variances in large samples. Accordingly, maximum likelihood estimation is the best approach for estimating the parameters in an ARMA model which is fitted to a given time series. Of particular, practical importance is the maximum likelihood estimator of McLeod (1977) described in Appendix A6.1 which is efficient both from statistical and computational viewpoints. This estimation procedure is used for estimating parameters in not only ARMA and ARIMA models but also many of the extensions to nonseasonal ARMA models presented later in the book and listed in Table 1.6.2.

Often the identification procedures of Section 5.3 suggest more than one model to fit to a specific time series. After calibrating the parameters for the ARMA models, the best overall model can be selected using the MAICE procedure of Section 6.3. The ways in which the MAICE approach can be incorporated into the three stages of model construction given in Figure III.I, are shown in Figure 6.3.1.

After selecting the best overall fitted model using the AIC or another appropriate ASC, the chosen model should be subjected to rigorous diagnostic checking. Procedures for making sure that various modelling assumptions are satisfied are described in detail in the next chapter.

APPENDIX A6.1

ESTIMATOR FOR ARMA MODELS

The purpose of this appendix is to describe the *modified sum of squares algorithm* of McLeod (1977) for obtaining approximate MLE's for the parameters in an ARMA model. As pointed out in Section 6.2.3 this estimator is computationally efficient and produces parameter estimates which are usually identical to the exact MLE's.

Let w_t, $t = 1,2,\ldots,n$, be a stationary time series which is normally distributed. One wishes to use the maximum likelihood estimator to obtain estimates for the parameters in the ARMA model defined in [3.4.4] or [4.3.4]. The parameters to estimate are:

1. the mean μ for the series. If the series has been differenced at least once, one may wish to set μ=0 for the w_t series. Otherwise, μ can be estimated using

$$\hat{\mu} = \bar{w} = \sum_{t=1}^{n} \frac{w_t}{n}$$

and then fixed at $\bar{w}$ when estimating the other model parameters. Another approach is to include μ as an additional parameter to estimate along with those mentioned below. For time series of moderate length (i.e., $n \geq 30$), the estimate given by $\bar{w}$ will be very close to that obtained when μ is iteratively estimated along with the other model parameters.

2. the p AR parameters contained in the set

$$\phi = (\phi_1, \phi_2, \ldots, \phi_p).$$

3. the q MA parameters in the set

$$\theta = (\theta_1, \theta_2, \ldots, \theta_q).$$

4. the innovation series given by $a_1, a_2, \ldots, a_n$.

5. the variance, σ_a^2, of the innovations.

To write down the likelihood function for an ARMA(p,q) model, one must assume a distribution for the innovations and hence the w_t series. In particular, assume the innovations are NID(0, σ_a^2) and the w_t sequence is $N(\mu, \sigma_w^2)$.

Recall that for a single random variable, w, which is $N(\mu, \sigma^2)$, the pdf is written as

$$p(w) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{\frac{-(w-\mu)^2}{2\sigma^2}\right\}$$

Suppose that one has a time series, w_t, of n random variables given by $w_1, w_2, \ldots, w_n$, where the w's are jointly normally distributed as

$$N(\mu, \Gamma_n^{(p,q)})$$

where

$$\mu^T = (\mu, \mu, \ldots, \mu)$$

since $\mu_1 = \mu_2 = \cdots = \mu_n$, $\Gamma_n^{(p,q)}(\phi, \theta) = E(\mathbf{w}\mathbf{w}^T)$ is the variance-covariance matrix of the w_t's where

$$\mathbf{w}^T = (w_1 - \mu, w_2 - \mu, \ldots, w_n - \mu)$$

and the (i,j) element of $\Gamma_n^{(p,q)}$ is $\gamma_{|i-j|}$ for which the autocovariance is defined by $\gamma_k = E[w_k, w_{t-k}]$ as in [2.5.3]. The joint normal distribution of the w_t's is given by

$$p(\mathbf{w}|\phi,\theta,\sigma_a^2,\mu) = (2\pi)^{-n/2}|\Gamma_n^{(p,q)}(\phi,\theta)|^{-1/2}\exp\left\{\frac{-\mathbf{w}^T(\Gamma_n^{(p,q)}(\phi,\theta))^{-1}\mathbf{w}}{2}\right\} \qquad [A6.1.1]$$

Let $\mathbf{M}_n^{(p,q)}(\phi,\theta) = \dfrac{\sigma_a^2}{\Gamma_n^{(p,q)}(\phi,\theta)}$ and, hence, $[\Gamma_n^{(p,q)}(\phi,\theta)]^{-1} = \dfrac{\mathbf{M}_n^{(p,q)}(\phi,\theta)}{\sigma_a^2}$. Then,

$$p(\mathbf{w}|\phi,\theta,\sigma_a^2,\mu) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_n^{(p,q)}(\phi,\theta)|^{1/2}\exp\left\{\frac{-\mathbf{w}^T\mathbf{M}_n^{(p,q)}(\phi,\theta)\mathbf{w}}{2\sigma_a^2}\right\} \qquad [A6.1.2]$$

When the unconditional sum of squares function is given by $S(\phi,\theta) = \sum_{t=-\infty}^{n} \hat{a}_t^2$, Box and Jenkins (1976, Ch. 7, A7.4) show that the above can be used to evaluate $\mathbf{w}^T\mathbf{M}_n^{(p,q)}(\phi,\theta)\mathbf{w}_n$ conveniently. The $\hat{a}_t$'s can be calculated using the back-forecasting procedure of Box and Jenkins (Box and Jenkins, 1976, pp. 215-220) as

$$\hat{a}_t = E[a_t|\mathbf{w},\phi,\theta]$$

More specifically, let $[w_t]$ and $[a_t]$ denote conditional expectations given $w_{t-1}, w_{t-2}, \ldots, w_1$. Then,

$$\phi(B)[w_t] = \theta(B)[a_t] \qquad [A6.1.3]$$

where $[a_t] = 0, t > n$. Similarly,

$$\phi(F)[w_t] = \theta(F)[e_t] \qquad [A6.1.4]$$

where F is the forward differencing operator defined by $Fw_t = w_{t+1}$, and $e_t \sim NID(0,\sigma_a^2)$ with $[e_t] = 0, t < 1$. Then, the unconditional sum of squares function $S(\phi,\theta)$ is calculated as follows:

Step 0: Initialization. Set Q large enough so the model is well approximated by a MA(Q) process. Typically, $Q \approx 100$ is sufficient.

Step 1: Calculate $[w_t]$ $(t = n+Q, \ldots, 1)$ using [A6.1.4]. Begin this calculation by setting $[w_t] = 0, t \geq n - p$.

Step 2: Calculate $[e_t]$ $(t = n+Q, \ldots, 1)$ using [A6.1.3]. Start this calculation by setting $[e_t] = 0$, $t = n - p$.

Step 3: Back forecast w_t $(t = 0, -1, \ldots, 1 - Q)$. This is done by using [A6.1.4] to calculate first $[w_o]$ then $[w_{-1}], \ldots, [w_{1-Q}]$.

Step 4: Calculate $[a_t]$ $(t = 1-Q, \ldots, n)$ using [A6.1.3].

Step 5: $S(\phi,\theta) = \sum_{t=1-Q}^{n} [a_t]^2$.

Consequently, the likelihood function is given by

$$L(\phi,\theta,\sigma_a^2|\mathbf{w}) \alpha \sigma_a^{-n} |\mathbf{M}_n^{(p,q)}(\phi,\theta)|^{1/2} \exp\left\{-\frac{S(\phi,\theta)}{2\sigma_a^2}\right\} \quad [A6.1.5]$$

Because the term $|\mathbf{M}_n^{(p,q)}(\phi,\theta)|$ is dominated by the expression $\exp\{-S(\phi,\theta)/2\sigma_a^2\}$ in [A6.1.5] for large n and $|\mathbf{M}_n^{(p,q)}(\phi,\theta)|$ is difficult to calculate, Box and Jenkins (1976, p. 213) suggest that the determinant can be disregarded and approximate MLE's can be obtained for the model parameters. However, if the sample is small and/or MA parameters are included in the model, the resulting parameter estimates may differ appreciably from the exact MLE's (McLeod, 1977). To rectify these problems various authors have suggested different approaches for calculating $|\mathbf{M}_n^{(p,q)}(\phi,\theta)|$. McLeod (1977) devised a procedure whereby $|\mathbf{M}_n^{(p,q)}(\phi,\theta)|$ is replaced by its asymptotic limit given by

$$m_{p,q}(\phi,\theta) = \lim_{n\to\infty} |\mathbf{M}_n^{(p,q)}(\phi,\theta)| \quad [A6.1.6]$$

When there are no MA parameters in an ARMA(p,q) model and hence $q = 0$, it is known that for $n \geq p$ (Box and Jenkins, 1976, p. 275)

$$|\mathbf{M}_n^{(p,0)}(\phi)| = |\mathbf{M}_p^{(p,0)}(\phi)|$$

and the matrix $\mathbf{M}_p^{(p,0)}(\phi)$ has the (i,j)th element (Pagano, 1973; McLeod, 1977)

$$\sum_{k=0}^{min(i,j)} (\phi_{i-k-1}\phi_{j-k-1} - \phi_{p+1+k-i}\phi_{p+1+k-j})$$

where $\phi_0 = -1$. To calculate $m_{p,0}(\phi)$, one can use

$$m_{p,0}(\phi,\theta) = |\mathbf{M}_p^{(p,0)}(\phi)|. \quad [A6.1.7]$$

As shown by McLeod (1977)

$$m_{p,q}(\phi,\theta) = \frac{m_{p,0}^2(\phi) m_{q,0}^2(\theta)}{m_{p+q,0}(\phi^*)} \quad [A6.1.8]$$

where ϕ^*_i is the ith parameter in the operator of order $p+q$ that is defined by

$$\phi^*(B) = \phi(B)\theta(B)$$

and $\phi^* = (\phi^*_1, \phi^*_2, \ldots, \phi^*_{p+q})$. Consequently, to compute $m_{p,q}(\phi,\theta)$ in [A6.1.8], it is only necessary to calculate the determinants of the three positive definite matrices which are obtained from [A6.1.7].

For convenience, McLeod (1977) defines the modified sum of squares function given by

$$S_m(\phi,\theta) = S(\phi,\theta)\{m_{p,q}(\phi,\theta)\}^{-1/n} \quad [A6.1.9]$$

and this is the function called the modified sum of squares (MSS) referred to in Section 6.2.3. To obtain MLE's for the model parameters, the modified sum of squares must be minimized by using a standard optimization algorithm such as the method of Powell (1964, 1965). When modelling seasonal time series, it is a straightforward procedure to appropriately alter [A6.1.9] for use with seasonal ARMA models (McLeod and Sales, 1983).

As noted in Section 6.1, often the value of the Box-Cox parameter λ in [3.4.30] is known in advance for a given type of time series. If λ is not known, this parameter can be iteratively estimated along with the other ARMA model parameters. However, one must take into account the Jacobian of the transformation to obtain the log likelihood function given by (McLeod, 1974; Hipel et al., 1977)

$$l(\lambda,\phi,\theta,\sigma_a^2) \approx -\frac{n}{2}\ln\frac{MSS}{n} + (\lambda - 1)\sum_{t=1}^{n}\ln(w_t + c) \qquad [A6.1.10]$$

where c is the constant in the Box-Cox transformation in [3.4.30] that causes all entries in the w_t series to be positive. When all the entries in the w_t series are greater than zero, one sets $c = 0$. When λ is fixed beforehand or estimated, one should minimize [A6.1.10] to obtain MLE's for the model parameters. If a computer package does not possess the capability of obtaining the MLE of λ, the log likelihood can be calculated for a range of fixed values of λ, and the λ which gives the largest value of the log likelihood can be chosen.

When using the estimator of this appendix to obtain MLE's for an ARMA model or other types of models given in this text, it is recommended that the w_t series be standardized before using the estimator. For example, each observation in the w_t series can be standardized by subtracting out the mean of the series and dividing this by the standard deviation of the series. If the series is not standardized, one may run into numerical problems when optimizing the likelihood function. This is especially true for the transfer function-noise and intervention models in Parts VII and VIII, respectively where the absolute magnitude of an estimated transfer function parameter may be much greater than the absolute magnitudes of the AR and MA parameters contained in the correlated noise terms.

APPENDIX A6.2

INFORMATION MATRIX

To obtain SE's for the MLE's of the AR and MA parameters in an ARMA model, one must calculate the variance-covariance matrix for the model parameters. The square roots of the diagonal entries in this matrix constitute the SE's for the corresponding parameter estimates.

Because the variance-covariance matrix is the inverse of the Fisher information matrix, first consider the definition for the *information matrix*. Let the sets of AR and MA parameters given in Section 6.2.1 as $\phi = (\phi_1,\phi_2, \ldots, \phi_p)$ and $\theta = (\theta_1,\theta_2, \ldots, \theta_q)$, respectively, be included in a single set as $\beta = (\phi,\theta)$. The variance of the innovations is denoted by σ_a^2. The likelihood function is written as $L(\beta|\mathbf{w})$, where $\mathbf{w} = (w_1,w_2, \ldots, w_n)$ is the set of observations. Let

$$I(\beta)=\left[\lim_{n\to\infty} E\left\{-\frac{\partial^2 \ln L(\beta|\mathbf{w})}{\partial\beta_i\partial\beta_j}|_{\beta=\beta}/n\right\}\right] \qquad [A6.2.1]$$

where the (i,j) element is defined inside the brackets on the right hand side, the dimension of the information matrix is $(p+q)$ by $(p+q)$, β_i and β_j are the ith and jth parameters, respectively, and $\hat{\beta}=(\hat{\phi},\hat{\theta})$ is the set of MLE's for the AR and MA parameters. Then, $I(\beta)$ is said to be the theoretical Fisher large sample information per observation on β. In practice $I(\beta)$ is estimated by $I(\hat{\beta})$.

The *variance-covariance matrix* for $V(\hat{\beta})$ for the set of MLE's $\hat{\beta}$ is given in large samples by the inverse of the information matrix. Hence,

$$V(\hat{\beta})\simeq nI(\hat{\beta})^{-1} \qquad [A6.2.2]$$

The square roots of the diagonal entries in the variance-covariance matrix in [A6.2.2] provide the estimates for the *standard errors (SE's)* of the corresponding parameters. The variance-covariance matrix is often referred to as simply the *covariance matrix*.

The second order partial derivatives with respect to the model parameters reflect the rate of change of slope of the log likelihood function. When this slope change is high, there is less spread around an optimum point in the log likelihood function. This in turn means that the inverse of the slope change is small which indicates a smaller SE when considering a diagonal entry in the variance-covariance matrix.

For an ARMA model, the variance-covariance matrix can be written in terms of the AR and MA parameters. In practice, the entries in the matrix can be calculated numerically.

Because it is known that MLE's are asymptotically normally distributed, one can test whether or not a given MLE is significantly different from zero. For example, if zero falls outside the interval given by the MLE ± 1.96 SE, one can state that the estimate under consideration is significantly different from zero at the 5% significance level. If this were not the case, one may wish to omit this parameter from the model fitted to the series. Constrained models are described in Section 3.4.4 while an example of a constrained model is the constrained AR(3) model fitted to the yearly St. Lawrence riverflow in Sections 5.4.2, 6.4.2 and 7.6.2.

From the definition [A6.2.1] it may be shown that

$$I(\beta)=\begin{bmatrix}\gamma_{vv}(i-j) & \gamma_{vu}(i-j)\\ \\ \gamma_{uv}(i-j) & \gamma_{uu}(i-j)\end{bmatrix} \qquad [A6.2.3]$$

where the (i,j) element in each partitioned matrix is indicated and $\gamma_{vv}(p\times p)$, $\gamma_{uu}(q\times q)$, $\gamma_{vu}(p\times q)$, $\gamma_{uv}(q\times p)$ are the theoretical auto and cross covariances defined by

$$\phi(B)v_t=-a_t,$$

$$\theta(B)u_t=a_t,$$

$$\gamma_{vv}(k) = E(v_t v_{t+k}),$$

$$\gamma_{uu}(k) = E(u_t u_{t+k}),$$

$$\gamma_{vu}(k) = E(v_t u_{t+k}),$$

$$\gamma_{uv}(k) = \gamma_{vu}(-k). \qquad [A6.2.4]$$

The covariance functions in [A6.2.4] may be obtained from a generalization of the algorithm given in Appendix A3.2.

APPENDIX A6.3

FINAL PREDICTION ERROR

Suppose that it is required to determine the order of an AR model to fit to a stationary time series $w_1, w_2, \ldots, w_n$. Prior to the introduction of the AIC defined in [6.3.1], Akaike (1969, 1970) developed a statistic called the *final prediction error (FPE)* for selecting the order of the AR model. The FPE is an estimate of the one step ahead prediction error variance of the AR(p) model in [3.2.5] and is defined as

$$FPE = \hat{\sigma}_a^2(p)\left(1 + \frac{p+1}{n}\right)\left(1 - \frac{p+1}{n}\right)^{-1} \qquad [A6.3.1]$$

where $\hat{\sigma}_a^2(p) = \frac{1}{n-p}\sum_{t=p+1}^{n} \hat{a}_t^2$ is the unbiased estimate of the residual variance of the AR(p) model. According to Akaike (1969, 1970), the AR model with the minimum value of the FPE in [A6.3.1] should be selected for modeling the series.

Taking natural logarithms of [A6.3.1] produces the result

$$\ln FPE = \ln\hat{\sigma}_a^2(p) + \frac{2(p+1)}{n} + 0(n^{-2}) \qquad [A6.3.2]$$

It is known that (-2) times the log likelihood of a Gaussian AR(p) model is approximately given by $n\ln\hat{\sigma}_a^2(p)$+ constant. Hence, as noted by Ozaki (1977),

$$n\ln FPE = AIC + constant + 0(n^{-1}) \qquad [A6.3.3]$$

Consequently, the MAICE procedure for AR model fitting is asymptotically equivalent to choosing the minimum value of the FPE.

PROBLEMS

6.1 Chapter 6 concentrates on explaining how the method of maximum likelihood can be used for estimating the parameters of ARMA models. However, other parameter estimation approaches are also available. Make a list of the names of six other estimation techniques. Outline the main ideas behind any two of these six methods.

6.2 In Section 6.2.2, first order and second order efficiency are referred to. Using equations where necessary, discuss these two concepts in more depth than that given in Section 6.2.2.

6.3 A criterion for characterizing an estimator is sufficiency. Define what is meant by sufficiency. Are maximum likelihood estimators sufficient?

6.4 What is an approximate maximum likelihood estimator? Outline the main components contained in the conditional and unconditional approximate maximum likelihood estimators suggested by Box and Jenkins (1976).

6.5 What is an exact maximum likelihoood estimator? Describe the main steps followed when applying the exact maximum likelihood estimators provided by Ansley (1979) as well as Ljung and Box (1979).

6.6 To optimize a likelihood or log likelihood function, a number of optimization algorithms are listed in Section 6.2.3. Outline the steps contained in the conjugate directions algorithm of Powell (1964, 1965). Discuss the advantages and limitations of Powell's algorithm.

6.7 Explain the difference between maximum likelihood estimation and Gaussian estimation.

6.8 Show that the exact log likelihood function for a Gaussian AR(1) is:

$$z_t = \mu + \phi_1(z_{t-1} - \mu) + a_t$$

where $a_t \sim NID(0,\sigma_a^2)$ and $t = 1, \ldots, n$ may be written as

$$\log L(\phi_1,\mu,\sigma_a^2) = -\frac{n}{2}\log\sigma_a^2 + \frac{1}{2}\log(1 - \phi_1^2) - \frac{1}{2\sigma_a^2}S(\phi_1)$$

where

$$S(\phi_1) = (1 - \phi_1^2)(z_1 - \mu)^2 + \sum_{t=2}^{n}[(z_t - \mu) - \phi_1(z_{t-k} - \mu)]^2$$

Simulate z_t, $t = 1, \ldots, n$ several times for various n and plot $L(\phi_1,\mu,\sigma_a^2)$ for $|\phi_1| < 1$.

6.9 Suppose that

$$(1 - \phi_1 B)z_t = a_t$$

where $\log a_t \sim NID(0,\sigma_a^2)$. Show that $I(\phi_1) = (1 + \sigma_a^{-2})e^{2\sigma_a^2}$

(a) Consider the AR(1) model

$$(1-\phi B)z_t = a_t$$

If $a_t \sim NID(0,1)$, show that $I(\phi) = \dfrac{1}{1-\phi^2}$

(b) Now consider the AR(1) model

$$(1-\phi B)z_t = a_t,$$

where $\log a_t \sim NID(0,1)$. Show that in this case, that

$$I(\phi) = 2e^2\left\{\frac{e(e-1)}{1-\phi^2} + \frac{e}{(1-\phi)^2}\right\}$$

(c) Compare the relative efficiency of Gaussian estimation versus maximum likelihood estimation when $\log a_t \sim NID(0,1)$. Verify your theoretical calculation by simulation (see Chapter 9 for an explanation of simulation).

6.10 Outline the theoretical development of the AIC given in [6.3.1].

6.11 Two approaches for employing the AIC in conjunction with model construction are described in Section 6.3.3. Using an annual time series of your choice, employ these two procedures for determining the best overall ARMA or ARIMA model to fit to the series.

6.12 Compare the MA(2) and AR(2) models for the Mean Annual Temperatures in Central England. Calculate the plausibility of the MA(2) model versus the AR(2) model.

6.13 The general form of an automatic selection criterion for model discrimination is given in Section 1.3.3 while the AIC and BIC are defined in [6.3.1] and [6.3.5]. Excluding the AIC and BIC, give the definitions of three other ASC's. Discuss the domains of applicability, advantages and drawbacks of each of these ASC's.

REFERENCES

AKAIKE INFORMATION CRITERION (AIC)

Akaike, H. (1971). Determination of the number of factors by an extended maximum likelihood principle. Research memorandum No. 44, The Institute of Statistical Mathematics, Tokyo.

Akaike, H. (1972a). Use of an information theoretic quantity for statistical model identification. In *Proceedings of the 5th Hawaii International Conference on Systems Sciences*, 249-250, Western Periodicals, North Hollywood, California.

Akaike, H. (1972b). Automatic data structure by maximum likelihood. In *Computers in Biomedicine*, Supplement to *Proceedings of 5th International Conference on Systems Sciences*, 99-101, Western Periodicals, North Hollywood, California.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csadki, F., Editors, *Proceedings of the 2nd International Symposium on Information Theory*, 267-281, Budapest. Akademial Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19:716-723.

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In Mehra, R. K. and Lainiotis, D. G., Editors, *System Identification,* 27-96. Academic Press, New York.

Akaike, H. (1978). On the likelihood of a time series model. Paper presented at the Institute of Statisticians 1978 Conference on Time Series Analysis and Forecasting, Cambridge University.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* 66(2):237-242.

Akaike, H. (1985). Prediction and entropy. In Atkinson, A. C. and Fineburg, F. E., Editors, *A Celebration of Statistics,* 1-24, Springer-Verlag, Berlin.

Hipel, K. W. (1981). Geophysical model discrimination using the Akaike information criterion. *IEEE Transactions on Automatic Control,* AC-26(2):358-378.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977a). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research,* 13(3):567-575.

Kitagawa, G. (1979). On the use of AIC for detection of outliers. *Technometrics,* 21(2):193-199.

McLeod, A. I. and Hipel, K. W. (1978). Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon. *Water Resources Research,* 14(3):491-508.

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, Applications. *Water Resources Research,* 13(3):577-586.

Ozaki, T. (1977). On the order determination of ARIMA models. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 26(3):290-301.

Sakamoto, Y. and Akaike, H. (1977). Analysis of cross classified data by AIC. *Annals of the Institute of Statistical Mathematics,* B:30-31.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika,* 63:117-126.

Tanabe, K. (1974). Fitting regression curves and surfaces by Akaike's Information Criterion. Research Memo, No. 63, The Institute of Statistical Mathematics, Tokyo.

Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability,* 12(3):488-497.

AUTOMATIC SELECTION CRITERIA BESIDES AIC

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics,* 21:243-247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics,* 22:203-217.

Akaike, H. (1977). On entropy maximization principle. In Krishnaiah, P. R., Editor, *Applications of Statistics,* 27-41. North-Holland, Amsterdam.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.

Chow, G. C. (1978). A reconcilliation of the information and posterior probability criteria for model selection. Research Memorandum No. 234, Econometric Research Program, Princeton University.

Gray, H. L., Kelley, G. D. and McIntire, D. D. (1978). A new approach to ARMA modelling. *Communications in Statistics*, B7(1):1-77.

Hannan, E. J. (1979). Estimating the dimension of a linear system. Unpublished manuscript, The Australian National University, Canberra, Australia.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Statistics*, 8:1071-1081.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, Series B, 41(2):190-195.

Hill, G. W. and Woodworth, D. (1980). Automatic Box-Jenkins forecasting. *Journal of the Operational Research Society*, 31(5):413-422.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661-675.

McClave, J. T. (1975). Subset autoregression. *Technometrics*, 17(2):213-220.

McClave, J. T. (1978). Estimating the order of autoregressive models: The Max χ^2 method. *Journal of the American Statistical Association*, 73(363):122-128.

Parzen, E. (1974). Some recent advances in time series modelling. *IEEE Transactions on Automatic Control*, AC-19(6):723-730.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, 14:465-471.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, 46(6):1273-1291.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461-464.

Shibata, R. (1989). Statistical aspects of model selection. Working Paper WP-89-077, International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society*, Series B, 41(2):276-278.

DATA SETS

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

ESTIMATION

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66:59-65.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society*, Series B, 11:115-149.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269-326.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cramer, H. (1946). *Mathematical Models of Statistics*. Princeton University Press.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, Cambridge, United Kingdom.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society*, Series A, 222:308-358.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700-725.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Hannan, E. J. (1970). *Multiple Time Series*. John Wiley, New York.

Kalman, R. E. (1960). A new approach of linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME*, Series D, 82:35-45.

Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*. The Iowa State University Press, Ames, Iowa.

Kendall, M. G. and Buckland, W. R. (1971). *A Dictionary of Statistical Terms*. Longman Group Limited, Thetford, Norfolk, Great Britain, third edition.

Kotz, S. and Johnson, N. L., Editors (1988). *Encyclopedia of Statistical Sciences, Volumes 1 to 9*. Wiley, New York.

Kruskal, W. H. and Tanur, J. M. (1978). *International Encyclopedia of Statistics, Volumes 1 and 2*. The Free Press, New York.

Li, W. K. and McLeod, A. I. (1988). ARMA modelling with non-Gaussian innovations. *Journal of Time Series Analysis*, 9(2):155-168.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. *Biometrika*, 66(2):265-270.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. *Biometrika*, 64(3):531-534.

McLeod, A. I. and Sales, P. R. H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of the Royal Statistical Society, Series C* (Applied Statistics), 32:211-223.

Mélard, G. (1984). Algorithm AS197. A fast algorithm for the exact likelihood of autoregressive-moving average models. *Journal of the Royal Statistical Society, Part C, Applied Statistics*, 33:104-114.

Mendel, J. M. (1987). *Lessons in Digital Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika*, 61(3):423-426.

Norden, R. H. (1972). A survey of maximum likelihood estimation. *International Statistical Review*, 40(3):329-354.

Norden, R. H. (1973). A survey of maximum likelihood estimation, part 2. *International Statistical Review*, 41(1):39-58.

Pagano, M. (1973). When is an autoregressive scheme stationary. *Communications in Statistics*, 1(6):533-544.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:531-546, University of California, Berkeley.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 24:46-72.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York, second edition.

Sachs, L. (1984). *Applied Statistics, A Handbook of Techniques*. Springer-Verlag, New York, second edition.

Schaerf, M. C. (1964). Estimation of the covariance autoregressive structure of a stationary time series. Technical report, Department of Statistics, Stanford University, Stanford, California.

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, 39:105-128.

OPTIMIZATION

Davidon, W. C. (1968). Variance algorithm for minization. *The Computer Journal*, 10:406-410.

Draper, N. R. and Smith, H. (1980). *Applied Regression Analysis*. Wiley, New York, second edition.

Gill, P., Murray, W. and Wright, M. (1981). *Practical Optimization*. Academic Press, New York.

Ishiguro, M. and Akaike, H. (1989). DALL: Davidon's algorithm for log likelihood maximization - a FORTRAN subroutine for statistical model builders. *The Institute of Statistical Mathematics*, Tokyo, Japan.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, Massachusetts, second edition.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society of Industrial and Applied Mathematics*, 11(2):431-441.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables with calculating derivatives. *Computer Journal*, 7:155-162.

Powell, M. J. D. (1965). A method for minimizing a sum of squares of nonlinear functions without calculating derivatives. *Computer Journal*, 8:303-307.

Vanderplaats, G. N. (1984). *Numerical Optimization Techniques for Engineering Design with Applications*. McGraw-Hill, New York.

CHAPTER 7

DIAGNOSTIC CHECKING

7.1 INTRODUCTION

In Chapter 5, a variety of useful graphical tools are presented for identifying one or more promising ARMA or ARIMA models to fit to a given time series. Subsequent to model identification, the method of maximum likelihood described in Chapter 6 can be employed for obtaining MLE's and SE's for the model parameters. When parameter estimates are calculated for more than one model, the AIC of Section 6.3, or another appropriate ASC mentioned in Section 6.3.6, can be used to select the overall best model. The objective of Chapter 7 is to ensure that this model adequately describes the time series under consideration by subjecting the calibrated model to a range of statistical tests which are referred to as *diagnostic checks*. The overall approach to model construction is displayed in Figure III.I while Figure 6.3.1 shows the ways in which the AIC can be used in conjunction with the model building stages.

One class of diagnostic checks is devised to test model adequacy by *overfitting*. This approach assumes that the possible types of model inadequacies are known in advance. The procedure of overfitting consists of including one or more extra parameters in the model to ascertain if an improved model can be designed (Box and Jenkins, 1976, Ch. 8; Granger and Newbold, 1977, Ch. 3). Section 7.2 explains how overfitting can be carried out in practice.

The most useful and informative diagnostic checks deal with determining whether or not the assumptions underlying the innovation series are satisfied by the residuals of the calibrated ARMA or ARIMA model. As pointed out in Section 3.4.5 and many other locations in the book, when fitting a model to a time series the estimated innovations or *residuals* are assumed to be independent, homoscedastic (i.e. have a constant variance) and normally distributed. Estimates for the a_t's are automatically calculated at the estimation stage along with MLE's and SE's for the model parameters (see Appendices A6.1 and A6.2).

Of the three innovation assumptions, *independence* and, hence, whiteness, is by far the most important. A data transformation cannot correct dependence of the residuals because the lack of independence indicates the present model is inadequate. Rather, the identification and estimation stages must be repeated in order to determine a suitable model. If the less important assumptions of *homoscedasticity* and *normality* are violated, they can often be corrected by a *Box-Cox transformation* of the data defined in [3.4.30].

Table 7.1.1 lists the main problems that can occur with the statistical properties of the residuals of a fitted model and how they can be corrected. Diagnostic checks for whiteness, normality and homoscedasticity of the residuals are presented in Sections 7.3 to 7.5, respectively, along with explanations regarding corrective actions that can be taken. Practical applications of applying these tests to a yearly riverflow series and sunspot numbers are presented in Section 7.6.

One should keep in mind that diagnostic checks only have meaning if the parameters of the model are efficiently estimated using the *maximum likelihood approach* of Chapter 6 at the estimation stage. If, for example, the method of moments were used to estimate the parameters of an ARMA model containing MA parameters, these moment estimates would be inefficient and

Table 7.1.1. Rectifying violations of the assumptions underlying the model residuals.

Violations of Residual Assumptions	Corrective Actions	Sections
Dependence and non-whiteness	Consider other models	7.3
Variance change or heteroscedasticity	Box-Cox data transformation	7.4
Non-normality	Box-Cox data transformation	7.5

probably quite different from the corresponding MLE's. Problems arising in the residuals of the ARMA model calibrated using moment estimates may be due to the inefficiency of the estimator rather than the specific parameters included in the model. Accordingly, for all of the diagnostic checks presented in Chapter 7 it is assumed that a maximum likelihood estimator is used to estimate the model parameters. For ARMA models, the only exception to this is the case of a pure AR model in [3.2.5]. Recall that for an AR model, both the method of moments using the *Yule-Walker equations* in [3.2.12] and the technique of maximum likelihood furnish efficient parameter estimates.

7.2 OVERFITTING

Overfitting involves fitting a more elaborate model than the one estimated to see if including one or more additional parameters greatly improves the fit. Extra parameters should be estimated for the more complex model only where it is feared that the simpler model may require more parameters. For example, the sample PACF and the IACF for an annual time series may possess decreasing but significant values at lags 1, 2, and 9. If an AR(2) model were originally estimated, then a model to check by overfitting the model would be

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_9 B^9)(w_t - \mu) = a_t$$

In Section 6.4.3, this is the type of model which is fitted to the square roots of the yearly sunspot numbers. Because, as shown in Table 6.4.3, the MLE of ϕ_9 is more than three times the value of its SE, this indicates that the more elaborate AR model containing ϕ_9 should be selected. Moreover, the AIC (Table 6.4.2) and diagnostic checks applied to the residuals of the constrained AR(9) model fitted to the square roots of the annual sunspot numbers (Section 7.6.3) confirm that the more complex model should be employed. Box and Newbold (1971, Section 3.6) as well as Box and Jenkins (1976, Section 8.1.2) show other interesting applications of overfitting.

The practitioner must take care to avoid *model redundancy* which could occur if the AR and MA components were simultaneously enlarged. For example, suppose that one initially fits an AR(1) model to a series but then expands the model by adding one more AR plus an additional MA parameter to form an ARMA(2,1) model. Suppose that the difference equation for an ARMA(2,1) model fitted to a given series given as w_t is

$$(1-0.80B+0.12B^2)(w_t-26)=(1-0.20B)a_t$$

Upon examining the SE's for some of the MLE's for the parameters one sees that they are very large. For example, the SE for ϕ_2 may be 0.22 which is much larger than $\hat{\phi}_2=0.12$, even though there are 200 entries in the series. The reason for a large SE is the instability introduced into the estimation algorithm due to parameter redundancy. Notice that the difference equation can be written as

$$(1-0.60B)(1-0.20B)(w_t-26)=(1-0.20B)a_t$$

which simplifies to

$$(1-0.60B)(w_t-26)=a_t$$

Therefore, the AR(1) model is more appropriate than the ARMA(2,1) model for fitting to the series.

Whenever one notices abnormally large SE's one should check for redundant or nearly redundant factors in a model due to overspecifying the model and then take corrective action by removing the redundant factors and fitting a simpler model. The over specification of the model parameters may cause rather large flat regions near the maximum point of the likelihood function and this in turn means that the SE's must be large (see Appendix A6.2). The large SE's suggest that a wide range of models could suitably model the data. However, in keeping with the principle of model parsimony, the simpler model should be chosen and, hence, redundancy should be avoided.

The problem of model redundancy provides an explanation as to why one cannot start out by fitting an overspecified model having many parameters and then reducing the number of parameters until an adequate model is found. Rather, one must begin with a fairly simple model and then carefully expand to a more complicated model, if necessary.

Another method of testing model adequacy by overfitting, which was originally suggested by Whittle (1952), is to fit a high-order AR model of order r where $20<r<30$. Suppose the original model has k estimated parameters plus the estimated residual variance, $\hat{\sigma}_a^2(k)$. Then it is shown (McLeod, 1974; Hipel et al., 1977) that the *likelihood ratio statistic* is

$$n\ln\left[\hat{\sigma}_a^2(k)/\hat{\sigma}_a^2(r)\right]\approx\chi^2(r-k) \qquad [7.2.1]$$

where $\hat{\sigma}_a^2(r)$ is the residual variance estimate for an AR process of order r. If the calculated $\chi^2(r-k)$ from [7.2.1] is greater than $\chi^2(r-k)$ from the tables at a chosen significance level, then a model with more parameters is needed.

The likelihood ratio test in [7.2.1] can also be used to determine if a model containing fewer parameters gives as good a fit as the full model. An application of this test is presented in Section 6.4.2 where three types of AR models are fitted to the average annual flows of the St. Lawrence River. The likelihood ratio test, as well as the AIC, select a constrained AR(3) without ϕ_2 as the best AR model to fit to the St. Lawrence flows.

When using the likelihood ratio test, the models being compared must be *nested*. Hence, the less complex model must be contained within the more complicated one. For instance, an AR(1) model is nested within an AR(k) model for $k\geq 2$. As pointed out in Section 6.3, when using the AIC for model discrimination, the models do not have to be nested and one can compare any number of different kinds of models at the same time.

7.3 WHITENESS TESTS

7.3.1 Introduction

The a_t sequence for AR (see Section 3.2), MA (Section 3.3), ARMA (Section 3.4) and ARIMA (Section 4.3) models are assumed to be independently distributed in the theoretical definition of these models. This implies that the estimated innovations or residuals are uncorrelated or white. In the next subsections, a number of statistical tests are described for determining whether or not the residuals, represented as $\hat{a}_t$, $t = 1,2, \ldots, n$, are white.

7.3.2 Graph of the Residual Autocorrelation Function

The most informative approach to check for whiteness is to examine a graph of the *residual autocorrelation function (RACF)*. The RACF at lag k is calculated as

$$r_k(\hat{a}) = \sum_{t=k+1}^{n} \left(\hat{a}_t \hat{a}_{t-k} / \sum_{i=1}^{n} \hat{a}_i^2 \right) \qquad [7.3.1]$$

Because of the term in the denominator in [7.3.1], the values of the RACF can range between -1 and +1. Additionally, since the RACF is symmetric about lag zero, one can plot the RACF against lags for positive lags from lag one to about lag $n/4$.

When examining a plot of the RACF, one would like to know if a given value is significantly different from zero. Asymptotically, the RACF is normally distributed as $N(0,\frac{1}{n})$ for any lag. Therefore, to draw the 95% confidence interval, for example, one can plot $\frac{+1.96}{\sqrt{n}}$ and $\frac{-1.96}{\sqrt{n}}$ above and below, respectively, the lag axis. If a given value of the RACF is significantly different from zero, it will fall outside the confidence interval.

A more accurate derivation for the large sample distribution of the RACF is provided by McLeod (1978). Define the vector of the first L values of the RACF as

$$\mathbf{r}(\hat{a}) = [r_1(\hat{a}), r_2(\hat{a}), \ldots, r_L(\hat{a})]' \qquad [7.3.2]$$

Denote by $\psi_k(\phi)$ the coefficient of B^k in the Maclaurin series expansion of $[\phi(B)]^{-1}$, where $\phi(B)$ is the AR operator defined in [3.4.4] as $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. Likewise, let $\psi_k(\theta)$ be the coefficient of B^k in the Maclaurin series expansion of $[\theta(B)]^{-1}$, where $\theta(B)$ is the MA operator given in [3.4.4] as $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$. Then it can be proven for large samples that the residuals in $\mathbf{r}(\hat{a})$ in [7.3.2] follow the multivariate normal distribution given as:

$$\mathbf{r}(\hat{a}) - N[\mathbf{0}, \frac{\mathbf{U}}{n}] \qquad [7.3.3]$$

where $\mathbf{U} = 1_L - X'I^{-1}X$, 1_L is the identity matrix, $I \approx X'X$ is the large-sample information matrix, and $X = [\psi_{i-j}(\phi), \psi_{i-j}(\theta)]$ are the i,j entries in the two partitions of the X matrix. The dimensions of the matrices X, $\psi_{i-j}(\phi)$, and $\psi_{i-j}(\theta)$ are, respectively, $L \times (p+q)$, $L \times p$, and $L \times q$.

Notice in [7.3.3] that U is a function of the AR and MA parameters in the ARMA model fitted to the original series. This is the reason why the findings are better than earlier work. Previously, Box and Pierce (1970) obtained [7.3.3] for an AR model but the result in [7.3.3] is valid for a more general ARMA model. Finally, equation [7.3.3] can be extended for use with seasonal ARIMA models (Section 12.3.4) as well as the other ARMA based models presented in Parts VI to IX.

To obtain the 95% confidence interval for the RACF at lag k, one calculates

$$\underset{\substack{\textit{confidence}\\\textit{interval}}}{95\%} = \pm 1.96\sqrt{\frac{1}{n}\mathrm{U}_{kk}}$$

where U_{kk} is the diagonal entry at location k,k in the matrix U in [7.3.3]. For each lag $k = 1,2,\ldots,L \simeq \frac{n}{4}$, one can determine the 95% confidence interval which can be plotted on a graph of the values of the RACF against lag k. Usually, the most important values of the RACF to examine are those located at the first few lags for nonseasonal data. If one or more of the values of the RACF fall outside the 95% confidence interval, this means that the current model is inadequate. The use of these confidence limits for checking model adequacy is discussed by Hipel et al. (1977), McLeod et al. (1977) and McLeod (1977).

When the present model is insufficient due to correlated residuals, one can use the results contained in a graph of the RACF to update the model. Suppose, for example, that an examination of the graph of the RACF reveals that the residuals of an AR(1) model fitted to the given w_t series are correlated at lag one. Hence, the inadequate model can be written as

$$(1 - \phi_1 B)(w_t - \hat{\mu}) = b_t$$

where ϕ_1 is the AR parameter, μ is the mean of the w_t series and b_t is the residual series that is correlated at lag one. Because the RACF has a significantly large value at lag one, the following MA(1) model can be fitted to the b_t series representing the correlated residuals:

$$b_t = (1 - \theta_1 B)a_t$$

where θ_1 is the MA parameter. By substituting b_t into the previous equation, one obtains the ARMA(1,1) model written as

$$(1 - \phi_1 B)(w_t - \mu) = (1 - \theta_1 B)a_t$$

Consequently, one can fit an ARMA(1,1) model to the original w_t series in order to obtain MLE's for the parameters when the parameters are all estimated together within the same ARMA(1,1) model framework. The residuals of the ARMA(1,1) can then be subjected to rigorous diagnostic checks in order to ascertain if further model modifications are required.

In the foregoing example for redesigning a model having correlated residuals, the form of the RACF clearly indicates how to expand the model. When this is not the case, other procedures can be employed for developing a more suitable model. One approach is to repeat the identification and estimation stages of model construction shown in Figure III.1 in order to discover a more suitable model. Another alternative is to use the AIC in conjunction with the earlier stages of model construction by following an appropriate path in Figure 6.3.1.

7.3.3 Portmanteau Tests

Rather than examine the magnitude of the value of RACF at each lag as is done in the previous subsection, one could look at an overall test statistic which is a function of the RACF values from lags one to L in order to perform a significance test for whiteness. However, this type of test is less sensitive because the lag locations of significantly large correlations and their magnitudes are buried in the test statistic. When a test statistic indicates a correlation problem in the RACF, one must then examine the graph of the RACF in order to understand what is happening and, subsequently, take corrective action.

Box and Pierce (1970) developed a Portmanteau statistic given as

$$Q'_L = n\sum_{k=1}^{L} r_k^2(\hat{a}) \qquad [7.3.4]$$

which is χ^2 distributed on $(L-p-q)$ degrees of freedom. Later, Davies et al. (1977) and Ljung and Box (1978) derived an improved version of the Portmanteau statistic which is written as

$$Q''_L = n(n+2)\sum_{k=1}^{L} r_k^2(\hat{a})/(n-k) \qquad [7.3.5]$$

and is also χ^2 distributed on $(L-p-q)$ degrees of freedom. More recently, Li and McLeod (1981) devised another enhanced Portmanteau statistic to test for whiteness. Specifically, if L is large enough so that the weights $\psi_k(\phi)$ and $\psi_k(\theta)$ in [7.3.3] have damped out, then

$$Q_L = n\sum_{k=1}^{L} r_k^2(\hat{a}) + \frac{L(L+1)}{2n} \qquad [7.3.6]$$

where Q_L is χ^2 distributed on $(L-p-q)$ degrees of freedom, and L can be given a value from about 15 to 25 for nonseasonal time series where L is not greater than about $n/4$. A test of this hypothesis can be done for model adequacy by choosing a level of significance and then comparing the value of the calculated χ^2 to the actual χ^2 value for $(L-p-q)$ degrees of freedom from the tables. If the calculated value is greater, on the basis of the available data the present model is inadequate, and appropriate changes must be made by examining in detail a plot of the RACF and, perhaps, also identification graphs of the original w_t series.

The modified Portmanteau statistics in [7.3.5] and [7.3.6] are recommended for employment over the first version in [7.3.4]. Moreover, the statistic in [7.3.6] has advantages over the one defined in [7.3.5]. In particular, using simulation experiments, Kheoh and McLeod (1992) demonstrate that the Portmanteau test statistic in [7.3.6] has a more accurate significance level than the one in [7.3.5] and possesses about the same power as that statistic. Also, the test statistic in [7.3.6] can be naturally extended for use in the multivariate case as in [21.3.2].

7.3.4 Other Whiteness Tests

A range of other whiteness tests can be employed for checking whether or not the residuals of a fitted ARMA model are white. For example, one can use the *cumulative periodogram graph* of Section 2.6 to test for whiteness. However, when examining model residuals, it is known that this test is inefficient. Often the cumulative periodogram test fails to indicate model inadequacy due to dependence of the residuals unless the model is a very poor fit to the given data.

A quite different approach to whiteness tests is to examine the *autocorrelation function (ACF) of the squared model residuals*, $\hat{a}_t^2$, $t = 1,2,\ldots,n$, which is estimated at lag k as

$$r_k(\hat{a}^2) = \sum_{t=k+1}^{n} \left[(\hat{a}_t^2 - \hat{\sigma}_a^2)(\hat{a}_{t-k}^2 - \hat{\sigma}_a^2)\right] / \left[\sum_{t=1}^{n}(\hat{a}_t^2 - \hat{\sigma}_a^2)^2\right] \qquad [7.3.7]$$

where the variance of the residuals is calculated using

$$\hat{\sigma}_a^2 = \sum_{t=1}^{n} \hat{a}_t^2 / n$$

Consider the vector of squared residuals given by

$$\mathbf{r}(\hat{a}^2) = [r_1(\hat{a}^2), r_2(\hat{a}^2), \ldots, r_L(\hat{a}^2)]^T \qquad [7.3.8]$$

For fixed L, McLeod and Li (1983) show that $\sqrt{n}\,\mathbf{r}(\hat{a}^2)$ is asymptotically multivariate normal with mean zero and unit covariance matrix. Hence, one could check for correlation of the squared residuals by examining a graph of $r_k(\hat{a}^2)$ against lag $k = 1,2,\ldots,L$, along with the 95% confidence limits. Furthermore, a significance test is provided by the Portmanteau statistic (Ljung and Box, 1978)

$$Q_L(\hat{a}^2) = n(n+2)\sum_{k=1}^{L} r_k^2(\hat{a}^2)/(n-k) \qquad [7.3.9]$$

which is asymptotically χ^2 distributed on $(L - p - q)$ degrees of freedom if the a_t are independent.

In some applications, the autocorrelation function of the squared residuals is more sensitive than the RACF for detecting residual dependence. In particular, the autocorrelation function of squared residuals have been found especially useful for detecting nonlinear types of statistical dependence in the residuals of fitted ARMA models (Granger and Andersen, 1978; Miller, 1979; McLeod and Li, 1983).

7.4 NORMALITY TESTS

7.4.1 Introduction

The theoretical definitions for AR, MA, ARMA and ARIMA models are presented in Sections 3.2.2, 3.3.2, 3.4.2, and 4.3.1, respectively. Recall that for each of these models it is assumed that the innovations, represented by the a_t's, are identically and independently distributed. This means that the disturbances must follow the same distribution, such as a Gamma or Gaussian distribution, and be independent of one another. As pointed out in Section 6.2, in order to obtain estimates for the model parameters one must assume that the innovations follow a specific distribution. In particular, comprehensive maximum likelihood estimators for ARMA models have been developed for the situation where the a_t's are Gaussian or normally distributed. A maximum likelihood estimator which is both statistically and computationally efficient is described in Appendix A6.1.

A wide range of flexible tests are available for ascertaining whether or not the residuals of a fitted ARMA model follow a normal distribution. Some of these normality tests are described in the subsequent subsections. If, for example, tests reveal that the residuals are not normal, one can transform the given data using the Box-Cox transformation in [3.4.30]. After fitting an ARMA model to the transformed series, one can employ appropriate normality tests to check whether or not the residuals from this model are Gaussian.

In addition to the statistical tests presented in the next three subsections and elsewhere, one can employ graphical methods for visually detecting departures from normality. A range of graphical techniques for use in exploratory data analysis are presented in Section 22.3 and referred to in Section 5.3.2. Some of these graphs can be used as visual normality checks. For example, if the box and whisker graph in Section 22.3.3 for the given time series is fairly symmetric, one can argue that the data follow a symmetric distribution such as a normal distribution. In a plot of the series against time, one should not see a lot of extreme values if the w_t series is Gaussian.

7.4.2 Skewness and Kurtosis Coefficients

Let the residual series for the fitted ARMA or ARIMA model be given as $\hat{a}_t$, $t = 1,2,\ldots,n$. If the $\hat{a}_t$'s are normally distributed, they should possess no significant skewness. The *skewness coefficient* g_1 for the $\hat{a}_t$ series is calculated using

$$g_1 = \left[\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^3\right] / \left[\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2\right]^{3/2} \qquad [7.4.1]$$

To test the null hypothesis that the data are normal and therefore possess no significant skewness, one must know the distribution of g_1. D'Agostino (1970) presents a method for transforming g_1 so that the transformed value is distributed as $N(0,1)$. This allows one to calculate the significant level for g_1.

The steps required in transforming g_1 to a random variable which is $N(0,1)$ are as follows (D'Agostino, 1970):

1. $Y = g_1\left[\frac{(n+1)(n+3)}{6(n-2)}\right]^{1/2}$ where g_1 is calculated from the $\hat{a}_t$ series using [7.4.1].

2. $B_2 = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$

3. $W^2 = -1 + [2(B_2 - 1)]^{1/2}$

4. $\delta = (\ln W)^{-1/2}$

5. $\alpha = [2/(W^2 - 1)]^{1/2}$

6. $$Z = \delta \ln[Y/\alpha + \left\{(Y/\alpha)^2 + 1\right\}^{1/2}]$$

The random variable Z, which is a transformation of the skewness coefficient g_1, is distributed as $N(0,1)$.

After calculating Z and choosing a level of significance, one can refer to standard normal tables to determine whether or not Z is significantly large. If, for example, Z has a significance level which is less than 0.05 according to the tables, one can assume that based upon the current information the $\hat{a}_t$ series possesses significant skewness and is, therefore, not normally distributed.

The *kurtosis coefficient* for the $\hat{a}_t$ series is determined as

$$g_2 = \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^4\right) / \left(\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2\right)^2 - 3 \qquad [7.4.2]$$

If the given data are normal, the statistic g_2 is approximately distributed as $N(0,24/n)$. Hence, for an estimated g_2, one can calculate the significance level for testing the null hypothesis that the data are normally distributed.

7.4.3 Normal Probability Plot

As before, suppose that a residual series is given as $\hat{a}_t$, $t = 1,2,\ldots,n$. When the entries in the $\hat{a}_t$ series are ordered from smallest to largest, the sample order statistic is

$$\hat{a}_{(1)} \leq \hat{a}_{(2)} \leq \cdots \leq \hat{a}_{(n)} \qquad [7.4.3]$$

Let the hypothesized cumulative distribution function of the transformed data be $F(\hat{a}/\hat{\sigma}_a)$. Also, let p_i, which is called the plotting position, be an estimate of $F(\hat{a}_{(i)}/\hat{\sigma}_a)$. Hence, $F^{-1}(p_i)$ is the theoretical standard quantile. To construct a probability plot, the $\hat{a}_{(i)}$ and $F^{-1}(p_i)$ are plotted as the abscissae and ordinates, respectively.

Following the recommendation of Looney and Gulledge (1985), for the case of a normal probability plot, the plotting position of Blom (1958) is recommended for use in practical applications. This plotting position is defined as

$$p_i = \frac{i - 0.375}{n + 0.25} \qquad [7.4.4]$$

When the $\hat{a}_t$'s are $N(0,\hat{\sigma}_a^2)$, a normal probability plot, consisting of the theoretical standard normal quantile $F^{-1}(p_i)$ being plotted against the empirical quantile $\hat{a}_{(i)}$, should form a straight line. The 95% Kilmogorov-Smirnov confidence interval (CI) can also be included with the normal probability plot. For a given plotting position, p_i, the two sides of the confidence interval are calculated using (Lilliefors, 1967)

$$95\%CI = \hat{a}_t + \hat{\sigma}_a \cdot F^{-1}\left(p_i \pm \frac{0.886}{\sqrt{n}}\right) \qquad [7.4.5]$$

The reader should keep in mind that this procedure is known to not be very sensitive to departures from normality, particularly in the tails. Additional research on probability plots includes contributions by Stirling (1982), Michael (1983) and Royston (1993).

7.4.4 Other Normality Tests

Besides those tests described in the previous two subsections, many other tests are available for determining whether or not a time series such as the sequence of model residuals is normally distributed. Normality tests are described in most standard statistical textbooks, statistical encyclopediae and handbooks, plus research papers. Shapiro et al. (1968), for instance, review and compare nine methods for testing for normality in a single sample. Two normality tests are briefly referred to below.

Shapiro-Wilk Test

The *Shapiro-Wilk test* for normality is based on the test statistic

$$W = b^2 / \sum_{t=1}^{n} \hat{a}_t^2 \qquad [7.4.6]$$

where b^2 is proportional to the best linear unbiased estimate of the slope of the linear regression of $\hat{a}_{(i)}$ in [7.4.3] on the expected value of the ith normal order statistic (Shapiro and Wilk, 1965). A general algorithm for calculating W and its significance level is given by Royston (1982). Simulation experiments suggest that the Shapiro-Wilk test is a good general omnibus test for normality in many situations. Finally, Filliben (1975) defines the normal probability plot correlation coefficient, which is closely related to the Shapiro-Wilk statistic, and compares the power of this test statistic for normality with six others.

Blom's Correlation Coefficient

Looney and Gulledge (1985) recommend the use of a correlation coefficient test for normality. The test, which is based upon Blom's plotting position, summarizes and objectively evaluates the information contained in a normal probability plot.

The test statistic for the composite test of normality is constructed using the Pearson product-moment correlation coefficient between $F^{-1}(p_i)$ and $\hat{a}_{(i)}$. As with the Shapiro-Wilk test, "large" values for the test statistic tend to support the assumption of normality. The significance range for the correlation coefficient test is obtained from the tabulated empirical percentage points printed in Looney and Gulledge's (1985) paper. Monte Carlo results indicate that this correlation coefficient test compares quite favourably to the Shapiro-Wilk test.

7.5 CONSTANT VARIANCE TESTS

7.5.1 Introduction

For the ARMA and ARIMA models of Chapters 3 and 4, respectively, as well as most of the other models in the book, the innovation series is assumed to have a constant variance, σ_a^2. The statistical word for constant variance is *homoscedasticity*. One would like the residuals of a fitted ARMA or ARIMA model to be homoscedastic.

If the variance of the innovations change, they are said to be *heteroscedastic*. Changing variance or heteroscedasticity can occur in a number of different ways. Firstly, the variance of the residuals may increase or decrease over time. Secondly, the variance may be a function of the magnitude of the series. For instance, the variance may be greater for higher values of the innovations and lower for smaller values. In the next section, tests are presented for checking for variance changes that occur over time and changes that are dependent upon level.

The plot of the Beveridge wheat price indices are shown in Figure 4.3.15. As can be seen, the variance or "spread" of the data is increasing over time. If an ARIMA model were fitted directly to the given time series, the variance of the residuals of the model would also become greater with increasing time. Consequently, as explained in Section 4.3.3, to alleviate problems with heteroscedasticity the wheat price indices are first transformed using the natural logarithmic transformation contained in [3.4.30] before fitting an ARIMA model to the series. In general, an appropriate Box-Cox transformation can often alleviate the problem of heteroscedasticity in the model residuals.

7.5.2 Tests for Homoscedasticity

The following tests were developed by McLeod (1974), and their application described by Hipel et al. (1977) and McLeod et al. (1977), are useful for determining whether a transformation of the data is needed by checking for changes in variance (heteroscedasticity) of the residuals. As is mentioned earlier, the variance of the normally independently distributed residuals is assumed to be constant (homoscedastic). Suppose that a_t is NID$[0,\sigma_a^2(t)]$ and that the variance changes with time as $\sigma_a^2(t)$. Let the stochastic random variable ζ_t be NID $(0,\sigma^2)$ and hence have constant variance. Suppose then that

$$a_t = \exp\left\{(\chi/2)[K(t) - \bar{K}]\right\}\zeta_t \qquad [7.5.1]$$

where χ is some constant to be estimated, $K(t)$ is a function of time to be specified, and $\bar{K}$ is the mean of $K(t)$ and equals $n^{-1}\sum_{t=1}^{n} K(t)$. The variance of the a_t residuals is then

$$\begin{aligned}\sigma_a^2(t) &= E\left\{\exp[\chi(K(t) - \bar{K})]\zeta_t^2\right\} \\ &= \exp\left\{\chi[K(t) - \bar{K}]\right\}\sigma^2 \qquad [7.5.2]\end{aligned}$$

It can be shown that the natural logarithm of the likelihood Lh for σ^2 and χ is

$$Lh = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left\{\exp[-\chi(K(t)-\bar{K})]a_t^2\right\} \qquad [7.5.3]$$

and

$$\frac{\partial Lh}{\partial\chi} = \frac{1}{\sigma^2}\sum_{t=1}^{n}\left\{K(t)\exp[-\chi(K(t)-\bar{K})]a_t^2\right\} \qquad [7.5.4]$$

One solves $\partial Lh/\partial\sigma^2 = 0$ exactly for σ^2, and substitutes for σ^2 into [7.5.4]. Next, equation [7.5.4] is set equal to zero, and the residual estimates $\hat{a}_t$ obtained from the estimation stage in Section 6.2 are used for a_t. This equation is then solved for a MLE of χ by using the Newton-Raphson method with an initial value of $\chi = 0$.

In order to carry out a test of the hypothesis, the first step is to postulate the null hypothesis that $\chi = 0$ and, therefore, to assume that the residuals have constant variance. The alternative hypothesis is that the residuals are heteroscedastic and that $\chi \neq 0$. By putting $K(t) = t$ in the previous equations, it is possible to test for trends in variance of the residuals over time. If $K(t) = w_t - \hat{a}_t$, then one can check for changes of variance depending on the current level of the w_t series in [4.3.3]. A likelihood ratio test of the null hypothesis is obtained by computing the MLE of χ and comparing it with its standard error. The variance for the MLE $\hat{\chi}$ for χ is calculated by using the equation

$$Var\hat{\chi} = -1/(\partial^2 Lh/\partial\chi^2) \qquad [7.5.5]$$

Because the MLE for χ is asymptotically normally distributed, after a level of significance is chosen it is a straightforward procedure to determine whether to accept or to reject the null hypothesis. This test is also valid for transfer function-noise, intervention, multivariate ARMA and regression models. In regression models, the test for heteroscedasticity can indicate whether an important covariate is missing (Anscombe, 1961; Pierce, 1971).

If model inadequacy is revealed by either of the tests, a simultaneous estimation procedure can be used to estimate the AR and MA parameters, σ^2, and χ. This would involve an enormous amount of computer time. However, in practice, the Box-Cox transformation in [3.4.30] will often stabilize the variance.

7.6 APPLICATIONS

7.6.1 Introduction

Tables 5.4.1 and 5.4.2 list ARMA and ARIMA models identified for fitting to five nonseasonal stationary and three yearly nonstationary time series, respectively. Detailed identification and estimation results are presented in Sections 5.4 and 6.4, respectively, for the average annual St. Lawrence riverflows and the yearly sunspot numbers. Likewise, in this section representative output from the diagnostic check stage of model construction is given for these same two annual geophysical time series. However, the reader should keep in mind that all of the models identified in Tables 5.4.1 and 5.4.2 passed the tests for whiteness, normality and homoscedasticity given in Sections 7.3 to 7.5, respectively.

7.6.2 Yearly St. Lawrence Riverflows

Figures 2.3.1 and 5.4.1 display the average annual flows of the St. Lawrence River (Yevjevich, 1963) in m^3/s at Ogdensburg, New York, from 1860 to 1957. Identification graphs in Figures 5.4.2 to 5.4.5 indicate that a constrained AR(3) model without ϕ_2 is the most appropriate AR model to fit to this series. Parameter estimates for this model along with their SE's are given in Table 6.4.1 while [6.4.2] is the difference equation for the calibrated model. Furthermore, both the likelihood ratio test (see [6.4.1] and [7.2.1]) and the AIC (see Section 6.3) select the constrained AR(3) model for describing to the St. Lawrence flows over the AR(1) and unconstrained AR(3) models, which are also listed in Table 6.4.1.

The St. Lawrence riverflow model in [6.4.2] is now subjected to rigorous diagnostic tests to ensure that the independence, normality and constant variance assumptions are satisfied. Figure 7.6.1 shows a plot of the RACF of Section 7.3.2 for the AR(3) model without ϕ_2. The 95% confidence limits in Figure 7.6.1 have jagged edges at low lags because the more accurate technique of Section 7.3.2, that is a function of both the fitted model parameters and the lag, is used to calculate these limits. Although the value of the RACF at lag 18 is rather large, it actually lies within the 1% significance interval. This larger value could be due to inherent random variation or to the length of the time series used to estimate it. However, the important values of the RACF for the lower lags all lie well within the 95% confidence interval. Therefore, the RACF indicates that the chosen model for the St. Lawrence River satisfies the whiteness assumption. This fact is also confirmed by the χ^2 distributed Portmanteau statistic Q_L in [7.3.6] whose calculated magnitude for Q_L is 13.46 for 18 degrees of freedom and is, therefore, not significant.

The less important assumptions of normality and homoscedasticity of the residuals are also satisfied. The skewness statistic g_1 in [7.4.1] has a value of -0.1482 and a SE of 0.3046. Because g_1 is much less than 1.96SE, there is no significant skewness and this indicates that the residuals are normally distributed. Likewise, the kurtosis coefficient in [7.4.2] confirms that the residuals are Gaussian. In particular, the kurtosis coefficient, g_2, has a value of -0.3240 which is less than its SE of 0.4974.

The χ statistic from Section 7.5.2 for changes in variance depending on the current level of the series has a magnitude of 0.000081 and a SE of 0.000341, while the χ statistic for trends in the variance over time possesses a value of 0.002917 with a corresponding SE of 0.00504. Because, in both instances, the SE's are greater than the χ statistics, based upon the information used, it can be assumed that the residuals are homoscedastic.

The flows used for the St. Lawrence River are in cubic meters per second. However, if the flows had been in cubic feet per second and a model had been fit to these data, all the AR parameters and SE's would have been identical with the metric model in [6.4.2]. Only the mean level of the series and $\hat{\sigma}_a^2$ would be different. In general, no matter what units of measurement are used the AR and the MA parameter estimates and the SE's will remain the same, while the mean level and $\hat{\sigma}_a^2$ will be different.

The type of model fit to the St. Lawrence River data reflects the actual physical situation. The Great Lakes all flow into the St. Lawrence River, and due to their immense size they are capable of over-year storage. If there is an unusually wet or an unusually dry year, the Great Lakes dampen the effect of extreme precipitation on the flows of the St. Lawrence River.

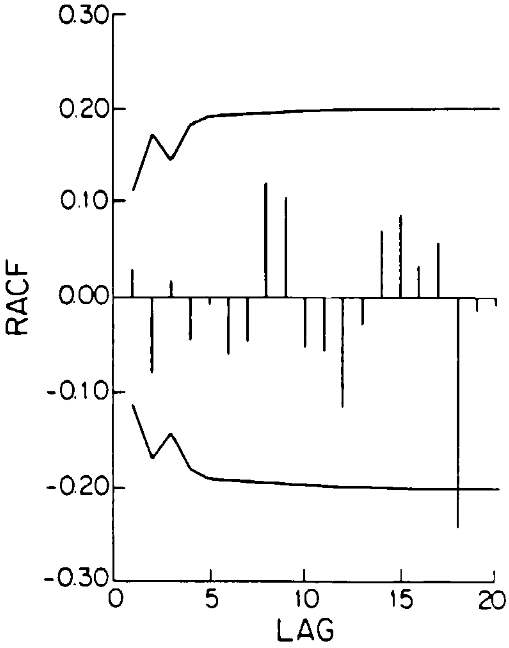


Figure 7.6.1. RACF and 95% confidence limits for the constrained AR(3) model without ϕ_2 fitted to the average annual flows of the St. Lawrence River from 1860 to 1957.

Because of this, the average annual flows are correlated, and the correct model is an AR process rather than white noise. For a general discussion of the employment of ARMA models in hydrology, the reader can refer to Section 3.6.

7.6.3 Annual Sunspot Numbers

Yearly Wolfer sunspot numbers are available from 1700 to 1960 (Waldmeier, 1961) and a plot of the series from 1770 to 1869 is shown in Figure 5.4.6. The identification graphs for this time series are presented in Figures 5.4.7 to 5.4.10. As explained in Section 5.4.3, these identification graphs in conjunction with diagnostic check output point out that an appropriate model to fit to the square roots of the sunspot series is a constrained AR(9) model without ϕ_3 to ϕ_8. In Section 6.4.3, the MAICE procedure also selects this model as the best overall ARMA model to describe the sunspot series. The finite difference equation for the best model is presented in [6.4.3] for the series of 100 sunspot values from 1770 to 1869 which is listed as Series E in Box and Jenkins (1976). In addition, the calibrated model for the entire sunspot series from 1700 to 1960 is written in [6.4.4].

The constrained AR(9) model in [6.4.4] without ϕ_3 to ϕ_8 satisfies all the modelling assumptions for the residuals. A plot of the RACF in Figure 7.6.2 shows that the residuals are uncorrelated. All of the estimated values of the RACF fall within the 5% significance interval. The χ^2 distributed portmanteau statistic Q_L in [7.3.9] has a value of 18.85 for 22 degrees of freedom. Therefore, the Q_L statistic in [7.3.6] also confirms that the residuals are not correlated. The diagnostic checks for homoscedasticity and normality of the residuals reveal that these assumptions are also fulfilled. The model in [6.4.4], therefore, adequately models the yearly Wolfer sunspot numbers. Other types of constrained models were examined, but the AR(9) process with ϕ_3 to ϕ_8 constrained to zero is the only model that is found to be satisfactory.

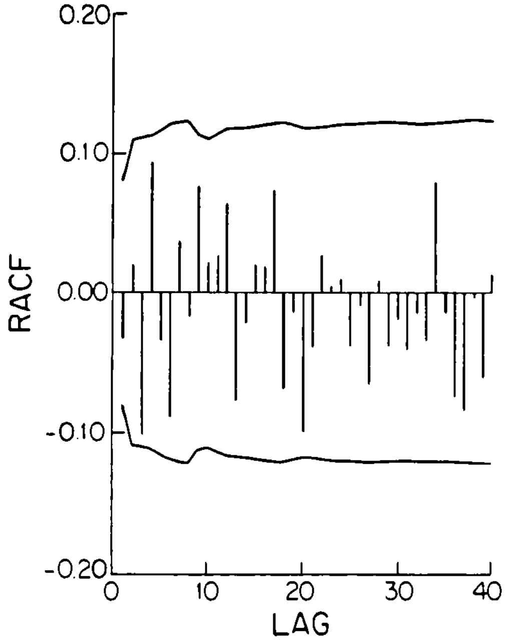


Figure 7.6.2. RACF and 95% confidence limits for the constrained AR(9) model without ϕ_3 to ϕ_8 fitted to the square roots of the yearly sunspot series from 1700 to 1960.

7.7 CONCLUSIONS

When fitting a time series model, such as an ARMA or ARIMA model, to a time series, one can follow the three stage procedure of model identification, estimation and diagnostic checking depicted in Figure III.I. The ways in which the AIC can enhance model construction are outlined in Figure 6.3.1. As explained in this and the previous two chapters, a variety of useful techniques are now available for allowing a practitioner to develop systematically and conveniently an appropriate model for describing a data set. The informative identification graphs of Section 5.3 permit a user to decide upon fairly quickly one or more tentative models to fit to the time series. These models can then be calibrated by using the method of maximum likelihood estimator presented in Appendix A6.1. When parameters for more than one model have been estimated, the AIC of Section 6.3 can be utilized to choose the overall best model. The model residuals can then be subjected to rigorous diagnostic checks to ascertain whether or not the residuals are white (Section 7.3), normally distributed (Section 7.4) and homoscedastic (Section 7.5). When the residuals are not white, then one must redesign the model by adding other parameters and, perhaps, eliminating unnecessary ones. The RACF of Section 7.3.2 is the best tool available for detecting nonwhiteness and assisting in developing a better model when the residuals are correlated. If residual problems are caused by non-normality and/or heteroscedasticity, these can often by corrected by invoking a Box-Cox transformation from [3.4.30] and then refitting the model.

The average annual riverflows of the St. Lawrence River at Ogdensburg, New York (Yevjevich, 1963), and the yearly sunspot numbers (Waldmeier, 1961) are used throughout Part III to explain clearly how model building is executed in practice. Some model building results are also referred to in Parts II and III for the other annual time series listed in Tables 5.4.1 and 5.4.2. For the case of the St. Lawrence riverflows, model identification plots in Figures 5.4.2 to 5.4.5 efficiently identify a constrained AR(3) model without ϕ_2 as being the best model to fit to the flows. In Section 6.4.2, the MAICE procedure and the likelihood ratio test confirm this as the most appropriate model to describe the series. Finally, the choice of a constrained AR(3) is reinforced by the diagnostic checks carried out in Section 7.6.2.

When examining the yearly sunspot numbers, the identification graphs of Figures 5.4.7 to 5.4.10 do not clearly pinpoint the most suitable ARMA type model to fit to the series. Rather, the need for a square root data transformation as well as the parameters required in the model are iteratively decided upon in Section 5.4.3 by examining a range of models. The final selection is a constrained AR(9) model without ϕ_3 to ϕ_8 that is fitted to the square roots of the sunspot numbers. In Section 6.4.3, the MAICE procedure also chooses this model from many possible candidates. When the constrained AR(9) model undergoes diagnostic testing for whiteness, normality and homoscedasticity in Section 7.6.3, the results confirm that the model is adequate.

After iteratively developing a model according to the steps in Figures III.1 and 6.3.1, one can use the calibrated model for practical applications. Two important applications of time series models are forecasting and simulation, which are now described in Part IV of the book.

PROBLEMS

7.1 Select an average annual time series that is of interest to you. Following the three stages of model construction and using an available time series program such as the MH Package mentioned in Section 1.7, fit the most appropriate ARMA(p,q) model to the data set. Overspecify the fitted model by adding an additional MA or AR parameter. Estimate the parameters of the overspecified model and comment upon the size of the SE's. Employ the likelihood ratio test of [7.2.1] to ascertain if overfitting is needed to start with and also to determine if the overfitted model is better than the simpler model.

7.2 An ARMA(2,1) model is written as

$$z_t - 0.13z_{t-1} + 0.36z_{t-2} = a_t - 0.4a_{t-1}$$

where it is assumed that the mean of z_t is zero. Can this model be written in a more parsimonious fashion?

7.3 Deliberately fit an overspecified ARMA or ARIMA model to an annual time series by assuming the model is ARMA(3,4). Comment upon the size of the SE's for the parameter estimates. Try to roughly factor this model to discover parameter redundancy. Determine the most appropriate model to fit to the time series.

7.4 Assume that one has an ARMA(1,1) model and $L = 5$ in [7.3.2] and [7.3.3]. Determine the entries of the matrix **U** in [7.3.3] for the distribution of the RACF.

7.5 Deliberately fit an underspecified ARMA or ARIMA model to a given annual time series. Based upon the RACF for this model, explain how the model can be expanded to provide a better fit to the series. If necessary, use other tools in your search for an improved model.

7.6 In Section 7.3.3, three versions of a Portmanteau statistic are presented for use in whiteness tests. By referring to appropriate references compare the relative advantages and drawbacks of the three statistics.

7.7 Explain why the autocorrelation function of the squared residuals is capable of detecting nonlinear statistical dependence in the residuals of fitted ARMA models.

7.8 The normality tests of Section 7.4 are described for use with the residual series from a fitted ARMA model. However, the tests can be employed with any given series such as the w_t series given in [4.3.3]. If the w_t series has a mean, then the mean should be subtracted from each w_t observation when calculating a given normality test statistic. Using a given annual series of your choice, determine if the series is Gaussian using the following tests:

(i) skewness coefficient,

(ii) kurtosis coefficient,

(iii) normality plot.

7.9 For a residual series obtained by fitting an ARMA model to a yearly time series, check for normality using the tests described in Sections 7.4.2 and 7.4.3.

7.10 Describe three additional normality tests beyond those given in Section 7.4.

7.11 A general test for homoscedasticity is described in Section 7.5.2. Assuming that one is checking for variance change over time and hence $K(t) = t$, describe in detail using equations how the test is carried out.

7.12 Select a yearly hydrological time series to model. Using a time series package, follow the three stages of model construction to ascertain the best ARMA or ARIMA model to fit to the data. Clearly explain all of your steps and show both identification and diagnostic check graphs.

7.13 In Figure 6.3.1, two main approaches are shown for using the AIC in model construction. Follow both of these approaches to find the most appropriate ARMA or ARIMA models for fitting to an annual riverflow series and also a yearly water demand series. Include both numerical and graphical results with your explanations of how you modelled the series.

REFERENCES

DATA SETS

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology paper no. 1, Colorado State University, Fort Collins, Colorado.

HOMOSCEDASTICITY TESTS

Anscombe, F. J. (1961). Examination of residuals. *In 4th Berkeley Symposium*, Berkeley, California.

Pierce, D. A. (1971). Distribution of residual autocorrelations in the regression model with autoregressive-moving average errors. *Journal of the Royal Statistical Society*, Series B, 33:140-146.

NORMALITY TESTS

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. John Wiley, New York.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika*, 57:679-681.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, 17(4):111-520.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399-402.

Looney, S. W. and Gulledge Jr., T. R. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician*, 39(1):75-79.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika*, 70(1):11-17.

Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115-124.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52:591-611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63:1343-1372.

Royston, P. (1993). Graphical detection of non-normality by using Michael's Statistic. *Applied Statistics*, 42(1):153-158.

Stirling, W. D. (1982). Enhancements to aid interpretation of probability plots. *The Statistician*, 31(3):211-220.

OVERFITTING

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. *Journal of the Royal Statistical Society, Series A*, 2:229-240.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press, New York.

Whittle, P. (1952). Tests of fit in time series. *Biometrika*, 39:309-318.

TIME SERIES ANALYSIS IN GENERAL

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research*, 13(3):567-575.

McLeod, A. I. (1974). Contributions to applied time series. Master's thesis, Department of Statistics, University of Waterloo, Waterloo, Ontario.

McLeod, A. I. (1977). *Topics in Time Series and Econometrics*. Ph.D. Thesis, Department of Statistics, University of Waterloo, Waterloo, Ontario, Canada, 283 pp.

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, Applications. *Water Resources Research*, 13(3):577-586.

WHITENESS TESTS

Box, G. E. P. and Pierce, D. A. (1970). Distribution of the residual autocorrelations in autoregressive integrated moving average models. *Journal of the American Statistical Association*, 65:1509-1526.

Davies, N., Triggs, C. M. and Newbold, P. (1977). Significance levels of the Box-Pierce Portmanteau statistics in finite samples. *Biometrika*, 64:517-522.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Gottingen.

Kheoh, T. S. and McLeod, A. I. (1992). Comparison of modified Portmanteau tests. *Journal of Computational Statistics and Data Analysis*, 14:99-106.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, Series B, 43(2):231-239.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:297-303.

McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society*, Series B, 40(3):296-302.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series*, 4(4):269-273.

Miller, R. B. (1979). Book review on "An introduction to bilinear time series models" by C. W. Granger and A. P. Anderson. *Journal of the American Statistical Association*, 74:927.

PART IV

FORECASTING

AND

SIMULATION

Within Part II of the book, two useful classes of nonseasonal models are defined and some of their theoretical properties are derived. In particular, the **ARMA** family of models of Chapter 3 are defined for fitting to stationary time series while the **ARIMA** class of models presented in Chapter 4 are designed for use with nonstationary data sequences. A sensible and systematic approach for fitting these and other kinds of models to a given data set is described in Part III. More specifically, by following the identification, estimation and diagnostic check stages of **model construction** explained in Chapters 5 to 7, respectively, one can develop the most appropriate model to describe the data set being studied.

A particular ARMA or ARIMA model which has been fitted to a time series can serve a variety of **useful purposes**. For example, the calibrated model provides an economic means of **encoding** the basic statistical properties of the time series into a few model parameters. In the process of carrying out the model building procedure, one obtains a **better understanding** about the key statistical characteristics of the data set. Besides the **insights** which are always gained when fitting a model to a time series, there are two important types of applications of time series models which are in widespread use by practitioners. These application areas are forecasting and simulation. The objectives of Part IV are to explain how ARMA and ARIMA models can be used for forecasting and simulation, and furnish case studies for demonstrating how forecasting and simulation are executed in practice.

The general purpose of **forecasting** or **prediction** is to provide the best estimates of what will happen at specified points in time in the future. Based upon the model fitted to a series and the most recent observations, one can obtain what are called **minimum mean square error forecasts** of future observations. Because forecasting is concerned with using the fitted model to extrapolate the time series into the future, it is often called **extrapolation**. Moreover, since forecasting, prediction or extrapolation provides an estimate of the future behaviour of a system, it is essential in the **operation** and **control** of the system. For example, forecasts for a riverflow series could be used for deciding upon the long range operating rules of a large reservoir. Forecasting can also be employed for **model discrimination.** When models from a variety of different classes are fitted to time series, one can select the model which provides the most accurate forecasts. The theory and practice of forecasting with nonseasonal models are presented in Chapter 8.

The overall objective of **simulation** is to use a fitted model to generate possible future values of a time series. These simulated or **synthetic sequences** can be used in two main ways. Firstly, simulated sequences can be utilized in **engineering design**. For instance, when designing a reservoir complex for generating hydroelectrical power, one can use both the historical flows and simulated data for obtaining the most economical design. Simulated sequences are employed in the design process because when the reservoir comes into operation the future flows will never be exactly the same as the historical flows. Therefore, one wishes to subject tentative designs to a wide variety of stochastically possible flow scenarios. Secondly, simulation can be employed for studying the **theoretical properties** of a given model. In many cases, it is very difficult or, for practical purposes, impossible, to determine precise analytical results for a given theoretical property of the model. When this is the situation, simulation can be used for obtaining the theoretical results to a specified desired level of accuracy. The theory and practice of simulating with nonseasonal models are explained in Chapter 9.

The forecasting and simulation techniques presented in the next two chapters are explained in terms of ARMA and ARIMA models. However, these methods can be easily extended for use with other models such as the different seasonal models of Part VI and the transfer function-noise models of Part VII. Table 1.6.3 lists the locations in the book where contributions to forecasting and simulation are given for a wide range of time series models.

CHAPTER 8

FORECASTING WITH NONSEASONAL MODELS

8.1 INTRODUCTION

In the design, planning and operation of water resources systems, one often needs good estimates of the future behaviour of key hydrological variables. For example, when operating a reservoir to serve multiple purposes such as hydroelectrical power generation, recreational uses and dilution of pollution downstream, one may require forecasts of the projected flows for upcoming time periods. The objective of forecasting is to provide accurate predictions of what will happen in the future.

In practical applications, forecasts are calculated after the most appropriate time series model is fitted to a given sequence of observations. Figure III.1 summarizes how a model is developed for describing the time series by following the identification, estimation and diagnostic check stages of model construction. Figure 6.3.1 outlines how an automatic selection criterion such as the AIC (Akaike information criterion) can be utilized in model building. After obtaining a calibrated model, one can calculate forecasts for one or more time steps into the future. Figure 8.1.1 displays the overall procedure for obtaining forecasts. Notice that the original data set may be first transformed using an appropriate transformation such as the Box-Cox transformation in [3.4.30]. Whatever the case, subsequent to constructing a time series model to fit to the series by following the procedures of Part III, one can use the calibrated model and the most recent observations to produce forecasts in the transformed domain. If, for example, an original data set of annual riverflows were first transformed using natural logarithms, then the forecasts from the ARMA model fitted to the logarithmic data would be predictions of the logarithmic flows. As indicated in Figure 8.1.1, one would have to take some type of inverse data transformation of the forecasted flows in the transformed domain in order to obtain forecasts in the original domain. These forecasts could then be used for an application such as optimizing the operating rules of a reservoir.

When calculating forecasts, one would like to obtain the most accurate forecasts possible. However, the question arises as to how one quantifies this idea of accuracy. One useful criterion for defining accuracy is to use what is called minimum mean square error. The theoretical definition of what is meant by *minimum mean square error forecasts* and the method of calculating them for ARMA and ARIMA models are presented in the next section. In addition, the method for calculating confidence limits for the forecasts is described.

Forecasting can be used as an approach for *model discrimination*. A variety of time series models can be fitted to the first portion of one or more time series and then used to forecast the remaining observations. By comparing the accuracy of the forecasts from the models, one can determine which set of models forecasts the best. In Section 8.3, *forecasting experiments* are carried out for deciding upon the best types of models to use with yearly natural time series.

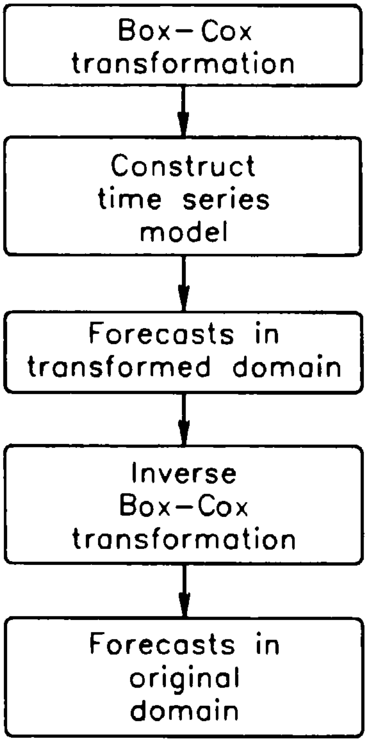

Box–Cox transformation
Construct time series model
Forecasts in transformed domain
Inverse Box–Cox transformation
Forecasts in original domain

Figure 8.1.1. Overall procedure for obtaining forecasts.

When comparing one step ahead forecasts, some statistical tests are described for determining if one model forecasts significantly better than another.

Chapter 8 deals with forecasting using nonseasonal ARMA and ARIMA models. Forecasting with other kinds of models is described in other chapters of the book. In Chapter 15, the procedures for forecasting with three types of *seasonal models* are presented. Forecasting experiments are also given in Sections 15.3 and 15.4 for comparing the forecasting abilities of different seasonal models. Procedures for combining forecasts from distinctly different models in order to obtain overall better forecasts are described in Section 15.5. Similar approaches could also be used for combining forecasts from different nonseasonal models. Finally, Chapter 18 describes how one can obtain forecasts using a *transfer function-noise model*. As explained in Chapter 17, a transfer function-noise model is a time series model that can describe situations where there is a single output and multiple inputs. For example, the output series may be riverflows whereas the input or covariate series are precipitation and temperature measurements. Table 1.6.3

summarizes where material on forecasting can be found in the book.

Besides hydrology, forecasting experiments have been carried out in other disciplines to compare the forecasting ability of models. In economics, one important forecasting study was completed by Newbold and Granger (1974). In their investigation these authors used one hundred and six economic time series to compare three types of forecasting models. The time series were split into two parts and ARIMA, Holt-Winters, and stepwise autoregressive models were fitted to the first portion of the data. The three models were then used to forecast the remainder of the data for various lead times. The forecasting ability of the three models was judged on the basis of the mean squared error (MSE) of the forecasts. Newbold and Granger (1974) found that the ARIMA forecasting procedure clearly outperformed the other two methods for short lead times but the advantage decreased for increasing lead time.

Madridakis et al. (1982) reported on a recent forecasting competition. The forecasting ability of over twenty models was tested using 1001 time series. The time series were of different length, type (i.e., monthly, quarterly, and annual) and represented data ranging from small firms to nations. Different forecast horizons were considered and several criterion were employed to compare the forecasts from the various models. In general, no one specific model produces superior forecasts for all types of data considered. However, some improvement may be achieved if the forecaster selected certain classes of models for forecasting specific types of data.

Because of the great import of forecasting in water resources engineering as well as many other disciplines, there have been many research papers, conference proceedings and books written on forecasting. Most of the water resources and time series analysis books referred to in Chapter 1 of this book contain chapters on forecasting. The Hydrological Forecasting Symposium (International Association of Hydrological Sciences, 1980) held in Oxford, England, certainly confirms the usefulness of forecasting in hydrology. For forecasting in economics, readers may wish to refer to texts listed in the references under economic forecasting at the end of Chapter 1. Within this book, recent practical developments for forecasting in water resources engineering are presented.

8.2 MINIMUM MEAN SQUARE ERROR FORECASTS

8.2.1 Introduction

Let z_t represent a known value of a time series observed at time t. For convenience of explanation, assume for now that the data have not been transformed using an appropriate data transformation. In Section 8.2.7, it is explained how the Box-Cox transformation in [3.4.30] is taken into account when forecasting. As shown in Figure 8.2.1, suppose that the observations are known up until time t. Given an ARMA or ARIMA model that is fitted to the historical series up to time t, one wishes to use this model and the most recent observation to forecast the series at time $t+l$. Let the forecast for the unknown observation, z_{t+l}, be denoted by $\hat{z}_t(l)$, since one is at time t and would like to forecast l steps ahead. The time t is referred to as the *origin time* for the forecast while l is the *lead time* which could take on values of $l = 1,2,...$. Consequently, in Figure 8.2.1, the forecasts from origin t having lead times of $l = 1$, 2, and 3, are denoted by $\hat{z}_t(1)$, $\hat{z}_t(2)$, and $\hat{z}_t(3)$, respectively. The forecast, $\hat{z}_t(1)$, at lead time 1, is called the *one step ahead forecast* and is frequently used in forecasting experiments for discriminating among competing models.

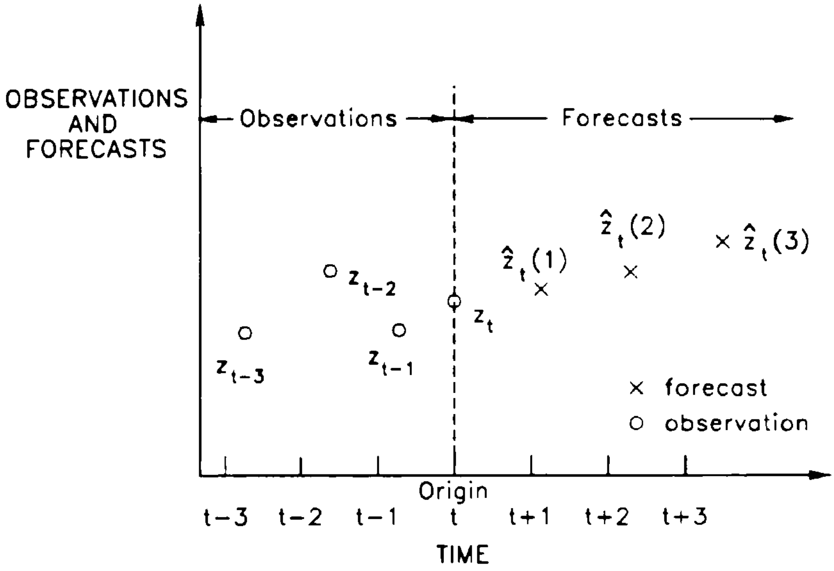


Figure 8.2.1. Forecasts from origin t.

One would like to produce forecasts which are as close as possible to what eventually takes place. Another way to state this is that one would like to minimize the forecast errors. This is because larger forecast errors can lead to poor decisions which in turn can cause more excessive costs than would be necessary. For example, if a hydroelectric complex were operated inefficiently because of poor forecasts, the utility could lose large sums of money.

To appreciate what is meant by *forecast errors*, refer once again to Figure 8.2.1. After the observation at time t+1 becomes known, the *one step ahead forecast error* from origin t is calculated as

$$e_t(1) = z_{t+1} - \hat{z}_t(1)$$

Likewise, the forecast errors for lead time two and three are determined, respectively, as

$$e_t(2) = z_{t+2} - \hat{z}_t(2)$$

$$e_t(3) = z_{t+3} - \hat{z}_t(3)$$

In general, the forecast error at lead time l is given as

$$e_t(l) = z_{t+l} - \hat{z}_t(l) \quad l = 1,2,... \qquad [8.2.1]$$

Decision makers would like to minimize forecast errors in order to keep the costs of their decisions as low as possible. However, when calculating forecasts for lead times $l = 1,2,\ldots,k$, how should one define the forecast error that should be minimized? For example, one approach is to minimize the *mean error* given by

$$\bar{e} = \frac{1}{k}\sum_{l=1}^{k} e_t(l) \qquad [8.2.2]$$

Another would be to minimize the *mean absolute error (MAE)* written as

$$MAE = \frac{1}{k}\sum_{l=1}^{k} |e_t(l)| \qquad [8.2.3]$$

A third alternative is to minimize the *mean square error (MSE)* defined as

$$MSE = \frac{1}{k}\sum_{l=1}^{k} e_t(l)^2 \qquad [8.2.4]$$

One could easily define other criteria for defining forecast errors to be minimized. For instance, one could weight the forecast errors according to their time distance from origin t and then use these weighted errors in any of the above types of overall errors. As pointed out in the next subsection, minimum mean square error forecasts possess many attractive properties that have encouraged their widespread usage in practical applications.

8.2.2 Definition

As explained in Sections 3.4.3 and 4.3.4 for ARMA and ARIMA models, respectively, these two classes of models can be written in any of the three equivalent forms:

1. difference equation form as originally defined,
2. random shock format (i.e. as pure MA model)
3. inverted form (i.e. as a pure AR model).

Any one of these three forms of the model can be used for calculating the type of forecasts defined in this section. However, for presenting the definition of what is meant by a *minimum mean square error (MMSE) forecast* when using an ARMA or ARIMA model the random shock model, is most convenient to use.

From [3.4.18] or [4.3.9], at time t, the random shock model is written as

$$\begin{aligned} z_t &= \psi(B)a_t \\ &= (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_t \\ &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \end{aligned} \qquad [8.2.5]$$

where $\psi(B) = (1 + \psi_1 B + \psi_2 B^2 + \cdots)$ is the random shock or infinite MA operator for which ψ_i is the ith parameter and a_t is the innovation sequence distributed as NID$(0,\sigma_a^2)$. When standing at time $t + l$, the random shock model is given as

$$\begin{aligned} z_{t+l} &= \psi(B)a_{t+l} \\ &= (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_{t+l} \\ &= a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots \end{aligned}$$

$$+ \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots \qquad [8.2.6]$$

For simplifying the explanation, the mean of the series is omitted in the above two equations and ensuing discussions. However, when there is a nonzero mean for z_t, all of the upcoming results remain exactly the same.

Suppose that standing at origin t, one would like to make a forecast $\hat{z}_t(l)$ of z_{t+l} which is a linear function of current and previous observations $z_t, z_{t-1}, z_{t-2}, \ldots,$. This in turn implies that the forecast is a linear function of current and previous innovations $a_t, a_{t-1}, a_{t-2}, \ldots,$. Using all of the information up to time t and the random shock form of the model in [8.2.6], let the best forecast at lead time l be written as

$$\hat{z}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots \qquad [8.2.7]$$

where the weights $\psi_l^*, \psi_{l+1}^*, \psi_{l+2}^*, \ldots,$ are to be determined. Notice that the innovations $a_{t+1}, a_{t+2}, \ldots, a_{t+l}$, and their corresponding coefficients are not included in [8.2.7] since they are unknown.

The theoretical definition of the mean square error of the forecast is defined as $E[z_{t+l} - \hat{z}_t(l)]^2$. By replacing z_{t+l} and $\hat{z}_t(l)$ by the expressions given in [8.2.6] and [8.2.7], respectively the mean square error is expanded as

$$\begin{aligned} E[z_{t+l} - \hat{z}_t(l)]^2 = E[&(a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots) \\ &- (\psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots)]^2 \end{aligned}$$

After expanding the right hand side by squaring and then taking the expected value of each term, the equation is greatly simplified because of the fact that

$$E[a_t a_{t-j}] = \begin{cases} 0, & j \neq 0 \\ \sigma_a^2, & j = 0 \end{cases}$$

More specifically, the equation reduces to

$$E[z_{t+l} - \hat{z}_t(l)]^2 = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 + \sum_{j=0}^{\infty} \left\{ \psi_{l+j} - \psi_{l+j}^* \right\}^2 \sigma_a^2 \qquad [8.2.8]$$

It can be seen that the above equation is minimized by setting $\psi_{l+j}^* = \psi_{l+j}$, $j = 0,1,2,\ldots,$ and thereby eliminating the second component on the right hand side of [8.2.8]. Consequently, when written in random shock form the MMSE forecast is derived as

$$\hat{z}_{t+l} = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots \qquad [8.2.9]$$

As noted by Box and Jenkins (1976, Ch. 5) the finding in [8.2.9] is a special case of more general results in prediction theory by Wold (1954), Kolmogorov (1939, 1941a,b), Wiener (1949) and Whittle (1963). By combining the result in [8.2.9] with the random shock model in [8.2.6]

$$\begin{aligned} z_{t+l} &= (a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots + \psi_{l-1} a_{t+1}) \\ &\quad + (\psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots) \\ &= e_t(l) + \hat{z}_t(l) \end{aligned} \qquad [8.2.10]$$

where $e_t(l)$ is the error of the MMSE forecast $\hat{z}_t(l)$.

8.2.3 Properties

The fact that [8.2.10] can be easily derived from the definition of a MMSE forecast, points out one of the advantages of using this kind of forecast. Fortunately, there are many other properties of a MMSE forecast that make it very beneficial for use in practical applications and some of them are described now. Let

$$E_t[z_{t+l}] = E[z_{t+l} | z_t, z_{t-1}, \cdots]$$

denote the *conditional expectation* of z_{t+l} given knowledge of all the observations up to time t. Then, attractive properties of a MMSE forecast include:

1. The MMSE forecast, $\hat{z}_t(l)$, is simply the conditional expectation of z_{t+l} at time t.

This can be verified by taking the conditional expectation of z_{t+l} in [8.2.10] to get

$$E_t[z_{t+l}] = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots = \hat{z}_t(l) \qquad [8.2.11]$$

Keep in mind that when deriving [8.2.11] the expression $E_t[a_{t+k}] = 0$ for $k > 0$, and $E_t[a_{t+k}] = a_{t+k}$ for $k \le 0$ since the innovations up to time t are known. Specific rules for calculating MMSE forecasts for any ARMA or ARIMA model are presented in the next subsection.

2. The *forecast error* is a simple expression for any ARMA or ARIMA model.

From [8.2.10], the forecast error from origin t and for lead time l is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \qquad [8.2.12]$$

3. One can conveniently calculate the *forecast error variance*.

In particular, the variance of the forecast error is

$$\begin{aligned} E[e_t(l)^2] &= V(l) = var[e_t(l)] \\ &= E[(a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots + \psi_{l-1} a_{t+1})^2] \\ &= (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \end{aligned} \qquad [8.2.13]$$

4. The MMSE forecast is *unbiased*.

This is because

$$E_t[e_t(l)] = E[a_{t+l} + \psi_1 a_{t+l-1} + \psi_2 a_{t+l-2} + \cdots + \psi_{l-1} a_{t+1}] = 0 \qquad [8.2.14]$$

where $E[a_{t+k}] = 0$ for $k > 0$.

5. The one step ahead forecast error is equal to the corresponding innovation and, therefore, one step ahead forecast errors are uncorrelated.

From [8.2.12], the one step ahead forecast error is

$$e_t(1) = z_{t+1} - \hat{z}_t(1) = a_{t+1} \qquad [8.2.15]$$

Because the innovations are independent, and, hence, uncorrelated, the one step ahead forecast errors must also be uncorrelated. As explained in Section 8.3, this result is very useful for developing tests to determine if one model forecasts significantly better than another.

6. Forecasts for lead times greater than one are, in general, correlated.

Consider forecasts for different lead times from the same origin t. Let the forecast errors for lead times l and $l+j$, where j is a positive integer, be given by $e_t(l)$ and $e_t(l+j)$, respectively. As shown by Box and Jenkins (1976, Appendix A5.1), the correlation between these two forecast errors is

$$corr[e_t(l),e_t(l+j)] = \frac{\sum_{i=0}^{l-1} \psi_i \psi_{j+i}}{\left\{\sum_{h=0}^{l-1} \psi_h^2 \sum_{g=0}^{l+j-1} \psi_g^2\right\}^{\frac{1}{2}}} \qquad [8.2.16]$$

Because of the correlation in [8.2.16], forecasts can lie either mainly above or below the actual observations when they become known.

7. Any linear function of the MMSE forecasts is also a MMSE forecast of the corresponding linear function of the future observations.

To explain what this means in practice, consider a simple example. Suppose that $\hat{z}_t(1)$, $\hat{z}_t(2)$, $\hat{z}_t(3)$ and $\hat{z}_t(4)$, are four MMSE forecasts. Then, $10\hat{z}_t(1) + 8\hat{z}_t(2) + 6\hat{z}_t(3) + 4\hat{z}_t(4)$ is a MMSE forecast of $10z_{t+1} + 8z_{t+2} + 6z_{t+3} + 4z_{t+4}$.

8.2.4 Calculation of Forecasts

Forecasting with ARMA Models

As explained in the previous subsection, the MMSE forecast, $\hat{z}_t(l)$, for lead time l, is simply the conditional expectation, $E_t[z_{t+l}]$, of z_{t+l} at origin t. When calculating the conditional expectations for an ARMA or ARIMA model one can write the model in any one of its three equivalent forms. These three formats are the difference equation form for the model as originally defined, random shock format and the inverted form (see Sections 3.4.3 and 4.3.4 for descriptions of the three forms for the ARMA and ARIMA family of models, respectively).

To simplify the notation required when determining MMSE forecasts, let the conditional expectations $E_t[a_{t+l}]$ and $E_t[z_{t+l}]$ be replaced by $[a_{t+l}]$ and $[z_{t+l}]$, respectively. For explaining how forecasts are determined using the three equivalent formats, consider the family of ARMA models defined in Chapter 3. For $l > 0$, the three equivalent formats for writing the MMSE forecasts are as follows:

Forecasts Using the Original Definition. By taking conditional expectations at time t of each term of the ARMA model in [3.4.3], the MMSE forecasts are:

$$[z_{t+l}] = \hat{z}_t(l) = \phi_1[z_{t+l-1}] + \phi_2[z_{t+l-2}] + \cdots + \phi_p[z_{t+l-p}] + [a_{t+l}]$$

$$-\theta_1[a_{t+l-1}] - \theta_2[a_{t+l-2}] - \cdots - \theta_q[a_{t+l-q}] \qquad [8.2.17]$$

As before, for convenience of explanation, the mean of the series is not written in the model. Following specific rules described below for calculating MMSE forecasts, one can easily determine each conditional expectation in [8.2.17].

Forecasts from the Random Shock Form. One can take conditional expectations at time t of the random shock form of the ARMA model in [3.4.18] to determine the MMSE forecasts as:

$$[z_{t+l}] = \hat{z}_t(l) = [a_{t+l}] + \psi_1[a_{t+l-1}] + \psi_2[a_{t+l-2}] + \cdots \qquad [8.2.18]$$

where ψ_i is the ith random shock parameter. When there are AR parameters in the original ARMA model, the number of innovation terms on the right hand side of [8.2.18] is infinite in extent. However, because the absolute values of the random shock parameters die off quickly for increasing lag, one can use a finite number of terms on the right hand side of [8.2.18] for calculating the forecasts up to any desired level of accuracy. Approaches for deciding upon how many MA parameters or terms to include in the random shock model are discussed in Section 3.4.3.

Forecasts using the Inverted Form. By taking conditional expectations at time t of the inverted form of the ARMA model in [3.4.25], the MMSE forecasts are:

$$[z_{t+l}] = \hat{z}_t(l) = [a_{t+l}] + \pi_1[z_{t+l-1}] + \pi_2[z_{t+l-2}] + \cdots \qquad [8.2.19]$$

where π_i is the ith inverted parameter. When there are MA parameters in the original model, the number of π_i parameters on the right side of [8.2.19] is infinite. Nonetheless, since the absolute values of the inverted parameters attenuate fairly quickly for increasing lag, only a finite number of inverted terms in [8.2.19] are required for calculating MMSE forecasts. Guidelines for deciding upon how many inverted components to include in the inverted form of the model are given in Section 3.4.3. In practice, only a moderate number of inverted parameters are needed.

Forecasting with an ARIMA Model

When forecasting with an ARIMA model, the simplest approach is to first calculate the *generalized nonseasonal AR operator* $\phi'(B)$ defined as

$$\phi'(B) = \phi(B)\nabla^d \qquad [8.2.20]$$

where $\phi(B)$ is the nonseasonal AR operator of order p, ∇^d is the nonseasonal differencing operator given in [4.3.3], and

$$\phi'(B) = 1 + \phi'_1B + \phi'_2B^2 + \cdots + \phi_{p+d}B^{p+d}$$

is the generalized nonseasonal AR operator for which ϕ'_i is the ith nonseasonal generalized AR parameter. The ARIMA model from [4.3.4] is then written as

$$\phi'(B)z_t = \theta(B)a_t \qquad [8.2.21]$$

where $\theta(B)$ is the nonseasonal MA operator of order q. By taking conditional expectations at time t of [8.2.21], the MMSE forecasts for an ARIMA model are determined using

$$[z_{t+l}] = \phi'_1[z_{t+l-1}] + \phi'_2[z_{t+l-2}] + \cdots + \phi'_{p+d}[z_{t+l-p-d}]$$
$$+ [a_{t+l}] - \theta_l[a_{t+l-l}] - \theta_2[a_{t+l-2}] - \cdots - [a_{t+l-q}] \qquad [8.2.22]$$

As pointed out in Section 4.3.1, usually the differenced series $w_t = \nabla^d z_t$ has a mean or level of zero. However, suppose this is not the case so that the model in [8.2.21] can be written as

$$\phi'(B)z_t = \theta_0 + \theta(B)a_t \qquad [8.2.23]$$

where the "deterministic component" $\theta_0 = \mu_w \phi'(1)$ and μ_w is the mean of the w_t series. For $d = 0$, 1 and 2 the term θ_0 can be interpreted as the level, slope in a linear deterministic trend, and quadratic trend coefficient, respectively. When the ARIMA model has the form of [8.2.23], forecasts are calculated recursively for $l = 1,2, \ldots,$ using

$$[z_{t+l}] = \theta_0 + \phi'_1[z_{t+l-1}] + \phi_2[z_{t+l-2}] + \cdots + \phi_{p+d}[z_{t+l-p-d}] + [a_{t+l}]$$
$$- \theta_1[a_{t+l-1}] - \theta_2[a_{t+l-2}] - \cdots - \theta_q[a_{t+l-q}] \qquad [8.2.24]$$

Rules for Forecasting

The most convenient equations to utilize when calculating MMSE forecasts are [8.2.17] and [8.2.22] for ARMA and ARIMA models, respectively. Whatever difference equation form of the ARMA or ARIMA model is employed for determining MMSE forecasts, one employs the simple rules listed below for the case of j being a non-negative integer to determine the conditional expectations written in these equations.

1. $$[z_{t-j}] = E_t[z_{t-j}] = z_{t-j}\,, \quad j = 0,1,2, \cdots \qquad [8.2.25]$$

   Because an observation at or before time t is known, the conditional expectation of this known value or constant is simply the observation itself.

2. $$[z_{t+j}] = E_t[z_{t+j}] = \hat{z}_t(j)\,, \quad j = 1,2, \cdots \qquad [8.2.26]$$

   The conditional expectation of a time series value after time t is the MMSE forecast that one wishes to calculate for lead time j from origin t.

3. $$[a_{t-j}] = E_t[a_{t-j}] = a_{t-j}\,, \quad j = 0,1,2, \cdots \qquad [8.2.27]$$

   Since an innovation at or before time t is known, the conditional expectation of this known value is the innovation itself. In practice, the innovations are not measured directly like the z_t's but are estimated when the ARMA or ARIMA model is fitted to the z_t or differenced series (see Chapter 6). Another way to determine a_t is to write [8.2.15] as

$$a_t = z_t - \hat{z}_{t-1}(1)$$

where $\hat{z}_{t-1}(1)$ is the one step ahead forecast from origin $t-1$.

4.

$$[a_{t+j}] = E_t[a_{t+j}] = 0\ , \quad j = 1,2, \cdots \qquad [8.2.28]$$

In the definition of the ARMA or ARIMA model, the a_t's are assumed to be independently distributed and have a mean of zero and variance of σ_a^2. Consequently, the expected value of the unknown a_t's after time t is zero because they have not yet taken place.

8.2.5 Examples

To explain clearly how one employs the rules from the previous section for calculating MMSE forecasts for both ARMA and ARIMA, two simple illustrative examples are presented. The first forecasting application is for a stationary ARMA model while the second one is for a nonstationary ARIMA model.

ARMA Forecasting Illustration

ARMA(1,1) models are often identified for fitting to annual hydrological and other kinds of natural time series. For example, in Table 5.4.1, an ARMA(1,1) model is selected at the identification stage for fitting to an annual tree ring series.

From Section 3.4.1 an ARMA(1,1) model is written in its original difference equation form for time $t+l$ as

$$(1 - \phi_1 B)z_{t+l} = (1 - \theta_1 B)a_{t+l}$$

or

$$z_{t+l} - \phi_1 z_{t+l-1} = a_{t+l} - \theta_1 a_{t+l-1}$$

or

$$z_{t+l} = \phi_1 z_{t+l-1} + a_{t+l} - \theta_1 a_{t+l-1}$$

By taking conditional expectations of each term in the above equation, the ARMA(1,1) version for [8.2.17] is

$$[z_{t+l}] = \phi_1[z_{t+l-1}] + [a_{t+l}] - \theta_1[a_{t+l-1}] \qquad [8.2.29]$$

Using the rules listed in [8.2.25] to [8.2.28], one can calculate the MMSE forecasts for various lead times l from origin t.

Lead Time l=1:

Substitute l = 1 into [8.2.29] to get

$$[z_{t+1}] = \phi_1[z_t] + [a_{t+1}] - \theta_1[a_t]$$

After applying the forecasting rules, the one step ahead forecast is

$$\hat{z}_t(1) = \phi_1 z_t + 0 - \theta_1 a_t = \phi_1 z_t - \theta_1 a_t$$

In the above equation, all of the parameters and variable values on the right hand side are known, so one can determine $\hat{z}_t(1)$.

Lead Time l=2:

After substituting $l = 2$ into [8.2.29], one obtains

$$[z_{t+2}] = \phi_1[z_{t+1}] + [a_{t+2}] - \theta_1[a_{t+1}]$$

Next one uses the rules from [8.2.25] to [8.2.28] to get

$$\hat{z}_t(2) = \phi_1\hat{z}_t(1) + 0 - \theta_1(0) = \phi_1\hat{z}_t(1)$$

where the one step ahead forecast is known from the previous step for lead time $l = 1$.

Lead Time $l \geq 2$:

When the lead time is greater than one, the forecasting rules are applied to [8.2.29] to get

$$\hat{z}_t(l) = \phi_1\hat{z}_t(l-1) + 0 - \theta_1(0) = \phi_1\hat{z}_t(l-1)$$

where the MMSE forecast $\hat{z}_t(l-1)$ is obtained from the previous iteration for which the lead time is $l - 1$.

ARIMA Forecasting Application

In Section 4.3.3, the most appropriate ARIMA model to fit to the total annual electricity consumption for the U.S. is an ARIMA(0,2,1) model. From Figure 4.3.10, one can see that the series is highly nonstationary and, therefore, differencing is required.

Following the general form of the ARIMA model defined in [4.3.3] and [4.3.4], the ARIMA(0,2,1) model is written at time $t+l$ as

$$(1 - B)^2 z_{t+l} = (1 - \theta_1 B)a_{t+l}$$

or

$$(1 - 2B + B^2)z_{t+l} = (1 - \theta_1 B)a_{t+l}$$

or

$$z_{t+l} - 2z_{t+l-1} + z_{t+l-2} = a_{t+l} - \theta_1 a_{t+l-1}$$

After taking conditional expectations of each term in the above equation, the forecasting equation is

$$[z_{t+l}] - 2[z_{t+l-1}] + [z_{t+l-2}] = [a_{t+l}] - \theta_1[a_{t+l-1}]$$

or

$$[z_{t+l}] = 2[z_{t+l-1}] - [z_{t+l-2}] + [a_{t+l}] - \theta_1[a_{t+l-1}] \qquad [8.2.30]$$

By employing the rules given in [8.2.25] to [8.2.30], one can determine the MMSE forecasts for lead times $l = 1,2, \cdots$, from origin t.

Lead Time l = 1:

Substitute $l = 1$ into [8.2.30] to obtain

$$[z_{t+1}] = 2[z_t] - [z_{t-1}] + [a_{t+1}] - \theta_1[a_t]$$

After invoking the forecasting rules, the one step ahead forecast is

$$\hat{z}_t(1) = 2z_t - z_{t-1} + 0 - \theta_1 a_t = 2z_t - z_{t-1} - \theta_1 a_t$$

Because all entries on the right hand side of the above equation are known, one can calculate $\hat{z}_t(1)$. Keep in mind that when fitting a model to a time series z_t, the historical z_t innovations are calculated at the estimation stage. Another way to calculate a_t is to write [8.2.15] as

$$a_t = z_t - \hat{z}_{t-1}(1)$$

where $\hat{z}_{t-1}(1)$ is the one step ahead forecast from origin $t-1$.

Lead Time l = 2:

After assigning $l = 2$ in [8.2.28], one gets

$$[z_{t+2}] = 2[z_{t+1}] - [z_t] + [a_{t+2}] - \theta_1[a_{t+1}]$$

In the next step, one uses the rules for calculating conditional expectations in order to obtain

$$\hat{z}_t(2) = 2\hat{z}_t(1) - z_t + 0 - \theta_1(0) = 2\hat{z}_t(1) - z_t$$

where the one step ahead forecast is determined in the previous iteration for which $l = 1$.

Lead Time l = 3:

Substitute $l = 3$ into [8.2.30] to obtain

$$[z_{t+3}] = 2[z_{t+2}] - [z_{t+1}] + [a_{t+3}] - \theta_1[a_{t+2}]$$

After applying the rules for calculating conditional expectations, the above equation becomes

$$\hat{z}_t(3) = 2\hat{z}_t(2) - \hat{z}_t(1) + 0 - \theta(0) = 2\hat{z}_t(2) - \hat{z}_t(1)$$

where the one and two step ahead forecasts from origin t are determined in the previous two iterations.

Lead Time $l \geq 3$:

When the lead time is greater than or equal to three, the forecasting rules are applied to [8.2.30] to obtain

$$\hat{z}_t(l) = 2\hat{z}_t(l-1) - \hat{z}_t(l-2) + 0 - \theta_1(0) = 2\hat{z}_t(l-1) - \hat{z}_t(l-2)$$

where the MMSE forecasts for $\hat{z}_t(l-1)$ and $\hat{z}_t(l-2)$ are determined in the two previous steps having lead times $l-1$ and $l-2$, respectively.

8.2.6 Updating Forecasts

When using the random shock form of the model, forecasts can be generated using [8.2.9] or [8.2.11]. The methods for calculating the random shock weights for ARMA and ARIMA models are presented in Sections 3.4.3 and 4.3.4, respectively. By using the random shock form of the forecasting model, one can develop an easy approach for efficiently updating forecasts. In particular, the forecasts $\hat{z}_{t+1}(l)$ and $\hat{z}_t(l+1)$ of the future observation z_{t+l+1} made from origins t+1 and t, respectively, are written following [8.2.11] as

$$\hat{z}_{t+1}(l) = \psi_l a_{t+1} + \psi_{l+1} a_t + \psi_{l+2} a_{t-1} + \cdots$$

$$\hat{z}_t(l+1) = \psi_{l+1} a_t + \psi_{l+2} a_{t-1} + \cdots$$

After subtracting the second equation from the first, one finds

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1} \qquad [8.2.31]$$

Because of this result, the forecast of z_{t+l+1} from origin t can be updated to become the forecast of z_{t+l+1} from origin t+1 by adding $\psi_l a_{t+1}$. From [8.2.15], one can see that a_{t+1} is simply the one step ahead forecast error from origin t.

In practice, the updating formula in [8.2.31] can be conveniently used for economizing on the number of computations for generating forecasts. Suppose one is at origin t and already has forecasts for lead times $l = 1,2, \ldots, L$. Immediately upon obtaining the next observation, z_{t+1}, one can calculate the forecast error $a_{t+1} = z_{t+1} - \hat{z}_t(1)$. This result can then be used to obtain forecasts $\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1}$ from origin t+1 for lead times $l = 1,2, \ldots, L-1$. Although the new forecast $\hat{z}_{t+1}(L)$ cannot be calculated using this method, it can be easily determined from the forecasts at shorter lead times using the original difference equation form of the model (see Section 8.2.4).

8.2.7 Inverse Box-Cox Transformations

The overall procedure for determining forecasts from a time series model is displayed in Figure 8.1.1. Before fitting a model to a given series, one may wish to transform the series using the Box-Cox transformation in [3.4.30] or some other appropriate transformation. As explained in Section 3.4.5, the purpose of the transformation is to rectify problems with non-normality and/or heteroscedasticity in the residuals of the fitted model. Whatever the case, when one uses the model constructed for the transformed series to obtain MMSE forecasts following the methods of Section 8.2.4, one determines forecasts in the transformed domain. For example, when an ARMA model is built for a logarithmic average annual riverflow series, the forecasts from the model are MMSE forecasts of the logarithmic flows. As pointed out in Figure 8.1.1, to get forecasts in the untransformed domain, one must take some type of inverse transformation of the forecasts.

There are two basic approaches for determining forecasts in the original units of the series being forecasted. The first procedure is to take the direct inverse transformation of the forecasts produced in the transformed domain. For instance, suppose that the original z_t series is transformed using natural logarithms. From [3.4.30], this transformation is written as

$$z_t^{(\lambda)} = \ln(z_t + c)$$

where the constant c is chosen just large enough to cause all of the entries in z_t to be nonnegative. For an average annual riverflow series, c would be equal to zero. Using the techniques of Section 8.2.4, one can obtain MMSE forecasts for $z_t^{(\lambda)}$ from origin t for any desired lead times. To get the forecasts in the untransformed domain, one can use the direct inverse logarithmic transformation written as

$$\tilde{z}_t(l) = \exp(\hat{z}_t^{(\lambda)}(l) - c) \qquad [8.2.32]$$

where $\hat{z}_t^{(\lambda)}(l)$, $l = 1,2,\ldots,$ is the MMSE forecast of $z_t^{(\lambda)}$ in the transformed domain and $\tilde{z}_t(l)$ is the corresponding forecast in the untransformed domain. The symbol for a MMSE forecast is not written above the forecast in the untransformed domain because usually the direct inverse transformation of a MMSE forecast in the transformed domain does not produce a MMSE forecast in the untransformed domain. When not using logarithms, the direct inverse Box-Cox transformation of the MMSE forecasts in transformed domain is written in the untransformed format as

$$\tilde{z}_t(l) = [\lambda\hat{z}_t(l) + 1]^{1/\lambda} - c \quad \text{where } \lambda \neq 0 \qquad [8.2.33]$$

Granger and Newbold (1976) call this the naive method since forecasts calculated using [8.2.37] or [8.2.38] are not the exact MMSE forecasts in the untransformed domain.

The second main approach for obtaining a forecast in the untransformed domain is to calculate the exact MMSE forecast (Granger and Newbold, 1976). More specifically, the exact MMSE forecast in the untransformed domain is determined from the fact that its transformed value follows a Normal distribution with expected value $\hat{z}_t^{(\lambda)}(l)$ and variance $V(l)$, where $V(l)$ is calculated using [8.2.13]. The expected value of the inverse Box-Cox transformed value is the desired MMSE forecast. Thus, the MMSE forecast, $\hat{z}_t(l)$, is given by

$$\hat{z}_t(l) = \frac{1}{\sqrt{2\pi V(l)}} \int_{-\infty}^{\infty} (\lambda y + 1)^{\frac{1}{\lambda}} e^{-\frac{1}{2}\frac{(y - \hat{z}_t^{(\lambda)}(l))^2}{V(l)}} dy, \quad \lambda \neq 0, \qquad [8.2.34]$$

and

$$\hat{z}_t(l) = e^{\hat{z}_t^{(\lambda)}(l) + \frac{1}{2}V(l)}, \quad \lambda = 0. \qquad [8.2.35]$$

The required integral in [8.2.34] may be determined numerically by Hermite polynomial integration.

In practice, it is found that the MMSE forecasts are slightly smaller than the corresponding naive forecasts. Also, studies with real data have shown that these minimum-mean-square-error forecasts do perform better than the naive forecasts.

8.2.8 Applications

Probability Limits

Models fitted to two annual time series are used for producing MMSE forecasts. In the first application, forecasts are calculated for an ARMA model describing a stationary series. The second forecasting example deals with forecasting using an ARIMA model fitted to a nonstationary series.

When plotting MMSE forecasts one should always include probability limits so that the variability in the forecasts can be properly appreciated. By using the formula for the variance of the forecast error in [8.2.13] and assuming normality one can calculate confidence limits. For example, the 50% probability limits for the l-step ahead MMSE forecast from origin t is

$$\hat{z}_t(l) \pm 0.674\sqrt{V(l)}$$

where $V(l)$ is the variance of the forecast error in [8.2.13]. When forecasting from origin t up to lead time L, one can calculate and plot the forecasts and 50% probability limits for $l = 1,2,\ldots,L$. Because the random shock parameters in [8.2.13] attenuate to zero for a stationary ARMA model, the forecasting probability limits asymptotically approach constant values for increasing l. On the other hand, the probability limits for forecasts from a nonstationary ARIMA model diverge for increasing l.

ARMA(1,1) Forecasts

A time series consisting of 700 tree ring indices from 1263 to 1962 is given by Stokes et al. (1973). The most appropriate ARMA model to fit to this series is the ARMA(1,1) model written in [3.4.15]. Following the rules given in Section 8.2.4, one can calculate MMSE forecasts for the calibrated tree ring model. Figure 8.2.2 displays the MMSE forecasts for lead times from 1 to 20. Notice that later observations in the series are plotted up to 1962. Starting from the origin 1962, MMSE forecasts are indicated from 1963 to 1982 along with their 50% and 90% probability intervals.

An example that explains how to calculate MMSE forecasts for an ARMA(1,1) model is given at the beginning of Section 8.2.5. Because the model is stationary, the forecasts for increasing lead times in Figure 8.2.2 draw closer to the mean of the series and the probability intervals run parallel to these forecasts. As would be expected, the best forecast for a future observation that is far from the last observation is the mean level.

ARIMA(0,2,1) Forecasts

Figure 4.3.10 portrays a graph of the total annual electricity consumption in the U.S.A. from 1920 to 1970 (United States Bureau of the Census, 1976). As explained in Section 4.3.3, the best ARIMA model to fit to this series is an ARIMA(0,2,1) model with an estimated Box-Cox transformation of $\lambda = 0.533$. Following the approach of Section 8.2.4, MMSE forecasts are first determined for the transformed domain where $\lambda = 0.533$. Subsequently, [8.2.35] is employed for calculating the MMSE forecasts shown in Figure 8.2.3 in the untransformed domain. An example of how to calculate MMSE forecasts by hand for an ARIMA(0,2,1) model without a Box-Cox transformation is given in Section 8.2.5.

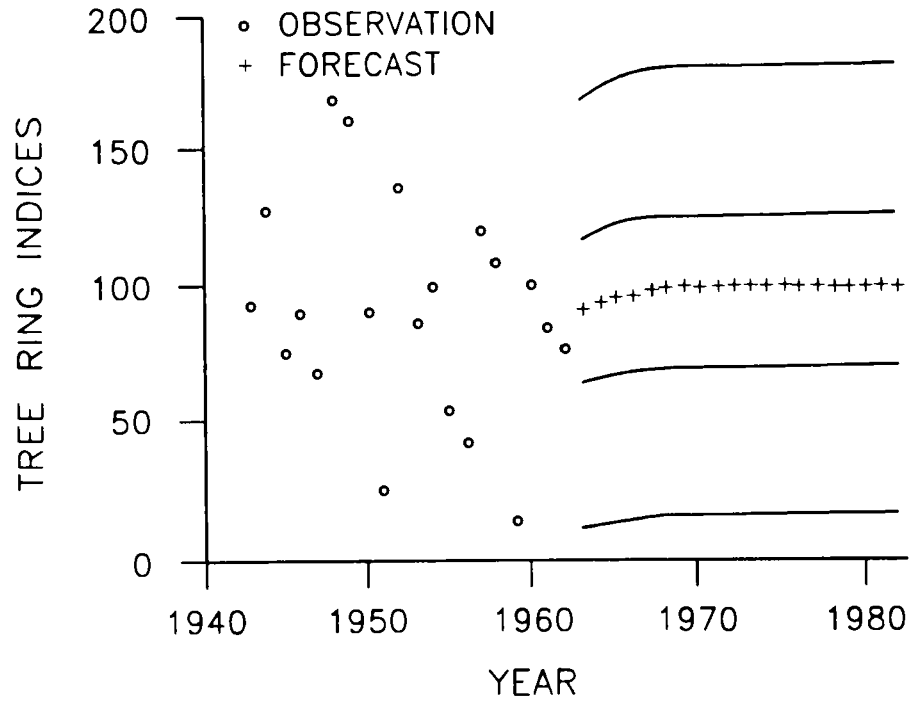


Figure 8.2.2. MMSE forecasts along with their 50% and 95% probability intervals for the ARMA(1,1) model fitted to the Douglas fir tree ring indices from Navajo National Monument, Arizona, U.S.A.

Figure 8.2.3 shows the MMSE forecasts in the untransformed domain calculated using the fitted model from 1971 to 1990 along with the 50% and 95% probability intervals. Because the series is nonstationary, observe how the forecasts continue the upward trend that is followed by the observations plotted on the left side of the figure. Moreover, the nonstationarity causes the probability limits to diverge outwards from the forecasts for increasing lead times.

8.3 FORECASTING EXPERIMENTS

8.3.1 Overview

An important test of the adequacy of a time series model is its ability to forecast well. The objective of this section is to employ forecasting experiments to demonstrate that ARMA models forecast very well when compared to other types of time series models that can be fitted to annual natural time series. This provides a sound reason for recommending the use of ARMA models by practitioners. In Sections 9.8 and 10.6, it is shown that ARMA models are also ideally suited for simulating hydrological as well as other types of natural phenomena.

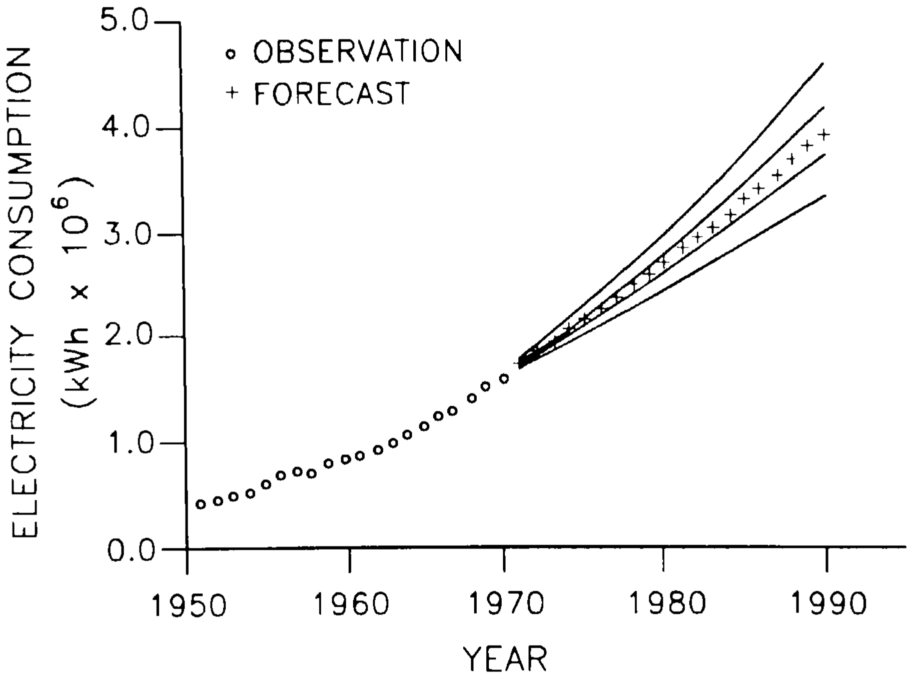


Figure 8.2.3. MMSE forecasts along with their 50% and 90% confidence intervals for the ARIMA(0,2,1) model fitted to the annual electricity consumption in the U.S.A.

In practical applications, *one step ahead forecasts* are often required for effectively operating a large-scale engineering project such as a system of reservoirs. When a new observation becomes available, the next one step ahead forecast can be made for deciding upon operating rules in the subsequent time period. Furthermore, a theoretical advantage of one step ahead forecasts is that they are statistically independent. This property allows one to develop statistical tests for determining if one model forecasts significantly better than another. In the next section, *statistical tests for comparing one step ahead forecasts* are presented and following this the different kinds of models used in the forecasting experiments are described.

To test the forecasting abilities of several stationary nonseasonal time series models, *split sample experiments* are performed in Section 8.3.4. Time series models are fitted to the first portion of the data in each of fourteen time series and these models are then employed to generate one step ahead forecasts. The forecasts errors are then compared using several loss functions to obtain ordinal rankings of the models. Statistical tests from Section 8.3.2 are then employed to test for significant differences in the forecasting performances of the various models. The forecasting results in the remaining part of Section 8.3 were originally presented by Noakes (1984) and Noakes et al. (1988).

8.3.2 Tests for Comparing Forecast Errors

Introduction

In the past, a great deal of effort has been devoted to the development of a wide variety of forecasting procedures. These procedures range from naive models or intuitive guesses to sophisticated techniques requiring skilled analysts and significant computer resources. At the same time, relatively little research has been devoted to developing methods for evaluating the relative accuracy of forecasts produced by the various forecasting methods.

In the forecasting experiments presented in Section 8.3.4, the forecast errors are examined from two different perspectives. Firstly, the performances of the various models are judged solely on the relative magnitudes of several criteria such as the *mean squared error (MSE)* or the *mean absolute percentage error (MAPE)* of the forecast errors. These comparisons provide ordinal rankings of the models but give no indication as to whether forecasts from a particular model are significantly better than forecasts from another model in a statistical sense. In order to address this question, a number of statistical tests are proposed to compare the performances of the models in a pairwise fashion and also to test the overall performances of particular models.

Wilcoxon Signed Rank Test

In order to ascertain whether the forecasts from a particular model are statistically significantly better than the forecasts generated by an alternative model, some form of statistical test must be employed. A nonparametric *Wilcoxon signed rank test* for paired data is one test which could be employed to test for significant differences in the forecasting ability of two procedures. This test was originally developed by Wilcoxon (1945) and is described in Appendix A23.2 in this book.

In this test, the differences in the squares of the forecast errors from two models for the same series are compared. These differences are ranked in ascending order, without regard to sign, and assigned ranks from one to the number of forecast errors available for comparison. The sum of the ranks of all positive differences is then computed as T in [A23.2.3] and compared to tabulated values in order to determine if the forecasts from a one model are significantly better than the forecasts from a competing model.

The results of this test may also be employed to examine the performances of the models across all of the series in the study. In this test, the probability associated with each T value is calculated by examining the area in the tail of the distribution. Fisher (1970, p. 99) presents a *combined level of significance test* such that

$$-2\sum_{i=1}^{k}\ln(p_i) \sim \chi^2_{2k} \qquad [8.3.1]$$

where p_i is the calculated probability associated with each T and k is the number of series considered in the test.

The Likelihood Ratio and Correlation Tests

It is of interest to examine statistically the difference in MSE's of the one step ahead predictor for two competing procedures in order to determine if the MSE's are significantly different. Thus, if $e_{1,t}$ and $e_{2,t}$ ($t = 1,2, \ldots, L$) denote the L one step ahead forecast errors for models 1 and 2 respectively, the null hypothesis is

$$H_0: MSE(e_{1,t}) = MSE(e_{2,t}) \qquad [8.3.2]$$

where $MSE(e) = <e^2>$ and $<.>$ denotes expectation. The alternative hypothesis, H_1, is the negation of H_0.

Granger and Newbold (1977, p. 281) have pointed out that a method originally developed by Pitman (1939) could be used to ascertain if one model forecasts significantly better than another. In this case, it is necessary to assume that $(e_{1,t}, e_{2,t})$ are jointly normally distributed with mean zero and are independent for successive values of t. In practice, the forecast errors may not be expected to satisfy all of the assumptions but these assumptions are probably a sensible first approximation. The assumptions of independence and zero mean seem quite reasonable if the forecasts are based on a good statistical model. As shown by Noakes (1984) and Noakes et al. (1988), a new test can be developed for the case in which the means are not known to be zero. For *Pitman's test*, let $S_t = e_{1,t} + e_{2,t}$ and $D_t = e_{1,t} - e_{2,t}$. Then Pitman's test is equivalent to testing if the correlation, r between S_t and D_t is significantly different from zero. Thus, provided $L > 25$, H_0 is significant at the five percent level if $|r| > 1.96/\sqrt{L}$. Previously, Pitman's test has often been used for testing the equality of variances of paired samples (Snedecor and Cochran, 1980, p. 190). It was pointed out in Lehmann (1959, p. 208, problem 33) that in this situation the test is unbiased and uniformly most powerful.

If the means of $e_{1,t}$ and $e_{2,t}$ are not both known to be zero, a likelihood ratio test can be employed. Let $(e_{1,t}, e_{2,t})$ be jointly normal with means (μ_1, μ_2) and covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

where σ_i^2 is the variance of the ith series and σ_{ij} is the covariance between i and j. Then the log likelihood for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$ is given by Rao (1973, p. 448) as

$$\log L(.) = \frac{L}{2}\log|(\sigma^{ij})| - \frac{1}{2}\sum\sum\sigma^{ij}[S_{ij} + L(\mu_i - \bar{\mu}_i)(\mu_j - \bar{\mu}_j)] \qquad [8.3.3]$$

where

$$\bar{\mu}_i = \frac{1}{L}\sum_{t=1}^{L} e_{i,t}$$

and

$$S_{ij} = L\sum_{t=1}^{L}(e_{i,t} - \bar{\mu}_i)(e_{j,t} - \bar{\mu}_j)$$

and $(\sigma^{ij}) = (\sigma_{ij})^{-1}$.

If L_0 is the maximized log likelihood assuming the null hypothesis is true and L_1 is the maximized log likelihood assuming the alternative hypothesis is true, then the *likelihood ratio statistic* is given by

$$R = 2(L_1 - L_0) \qquad [8.3.4]$$

When H_0 is true, it can be shown that $R \approx \chi_1^2$ (Rao, 1973).

If it is assumed that the means of the two error series are zero, then ignoring constants, the maximized log likelihoods are

$$L_1 = -\frac{L}{2}\log(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\sigma}_{12}^2) \qquad [8.3.5]$$

and

$$L_0 = -\frac{L}{2}\log(\hat{\sigma}^2 - \hat{\sigma}_{12}^2) \qquad [8.3.6]$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimated forecast error variances for the two competing models, $\hat{\sigma}_{12}$ is the estimated covariance of the estimated forecast errors and

$$\hat{\sigma}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2} \qquad [8.3.7]$$

The resulting likelihood ratio is then calculated using [8.3.4].

Equation [8.3.3] is easily maximized analytically when there are no restrictions on the parameters and so the maximized log likelihood is obtained. Under H_0

$$\sigma_1^2 + \mu_1^2 = \sigma_2^2 + \mu_2^2 \qquad [8.3.8]$$

and the log likelihood may be maximized numerically over $(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\sigma_{12})$ with $\sigma_2^2 = \sigma_1^2 + \mu_1^2 - \mu_2^2$. The conjugate direction minimization algorithm of Powell (1964) with a penalty function to ensure that $\sigma_2^2>0$ is recommended. Thus, the likelihood ratio test statistic, R, which is χ_1^2 under H_0 is obtained from [8.3.4].

8.3.3 Forecasting Models

Introduction

Stationary nonseasonal time series models are of particular interest to hydrologists since they often wish to model annual time series that are approximately stationary over a specified time period and subsequently use the fitted models for forecasting and simulation. Furthermore, stationary nonseasonal models form the foundations for seasonal (see Part VI) models as well as other kinds of models (see Parts VII to IX).

When fitting a nonseasonal stationary model, or for that matter any type of stochastic model, to a given data set, one can follow the identification, estimation and diagnostic check stages of model construction described in Part III as well as elsewhere in the book. Figure III.1 depicts this systems design approach to model building while Figure 6.3.1 shows how the AIC can enhance model construction. All of the different kinds of models employed in the forecasting studies are carefully developed following this sensible approach to model building.

The five families of stationary nonseasonal models used in the study are as follows:

1. ARMA (see Chapter 3, Part III, and Section 8.2 for definition, model building and forecasting, respectively),
2. FGN (Fractional Gaussian Noise, see Section 10.4.5),
3. FARMA (fractional ARMA, see Chapter 11) and FDIFF (fractioning differencing, special case of FARMA models in Chapter 11),
4. Markov (see this section),
5. Nonparametric (see this section).

Following the definition in Section 2.5.3, the second and third models have *long memory* while the remaining ones possess *short memory*. Additionally, the first three types of models are described at other indicated sections in this book while the last two are now outlined.

Markov and Nonparametric Regression Models

A number of researchers have proposed various nonparametric models for modelling and forecasting hydrological time series (see for example Denny et al. (1974) and Yakowitz (1973, 1976, 1979a,b, 1985a,b)). These models offer an attractive alternative to the ARMA as well as long memory FGN and FARMA models. The flexibility and modest computational requirements associated with nonparametric models are certainly two important considerations in model selection. As well, probability statements can be made concerning forecasted events. In light of these attractive characteristics, two nonparametric models are considered in this forecasting study.

A First Order Markov Model: The underlying concepts associated with stationary Markov chains are well known and explained in many standard statistical and operational research books. The first model considered is a first order Markov process defined as

$$Pr(X_{n+1}|X_n,X_{n-1},\cdots)=Pr(X_{n+1}|X_n) \qquad [8.3.9]$$

Although higher order processes may be required to adequately model the data, the first order approximation is a reasonable first step.

The time series data are first arranged in ascending order. If there are n data, m = integer $(\sqrt{n})$ states are selected at equal intervals. For example, if $n = 100$, then 10 states would be selected. The first 10 data would then be assigned to the first state and the state mean would be the arithmetic mean of these elements. This procedure is repeated until the m state means are calculated.

Based upon this arbitrary selection of states and estimated state means, each datum is reassigned to a specific state according to the Euclidean distance between the observation and the state means. That is, X_i is in state v if

$$|X_i - c_v| \le |X_i - c_k|, \quad 1 \le k \le m \qquad [8.3.10]$$

where the c_k's are the state means. A check is then made to ensure that at least $n^{1/3}$ data are associated with each state.

Quasi state transition probabilities are then estimated using the original time series and the selected states. Forecasts can then be made using these transition probabilities and the state means.

A Nonparametric Model: Yakowitz (1985a,b) employs nonparametric regression techniques to develop a more comprehensive and flexible nonparametric model. Unlike the simple first order Markov model outlined above, this nonparametric model allows for higher order dependence. A method for forecasting using this new model is also presented by Yakowitz (1985a,b).

Kernel nonparametric estimators of the density by Rosenblatt (1956, 1971) as well as kernel nonparametric regression estimators introduced by Watson (1964) have been extensively investigated and have also found practical application in fields such as pattern recognition. They can be briefly described as follows. Suppose that there are n independent observations, Y_i, $i = 1,2,\ldots,n$ with common density $f(y)$. Then the estimate of $f(y)$ based on the kernel $k(\cdot)$ is given by

$$\hat{f}(y) = \frac{1}{n\alpha_n}\sum_{i=1}^{n} k\left[\frac{y - Y_i}{\alpha_n}\right] \qquad [8.3.11]$$

where α_n is called a smoothing parameter and $k(\cdot)$ is generally taken to be a probability density function such as the standard normal. The choice of the kernel, $k(\cdot)$, is not as crucial as is the choice of the parameter α_n to obtain a good estimate.

For the regression case, suppose that one observes pairs of independent and identically distributed variables (Y_i,X_i) and that one wishes to estimate the expectation of $g(Y)$ conditional on the value $X = x$, where the pair (Y,X) has the same distribution as the observations (Y_i,X_i), $i = 1,2,\ldots,n$, and $g(\cdot)$ is a real function. The estimate of $E[g(Y)|X = x]$ is given by (Watson, 1964)

$$\hat{E}[g(Y)|X = x] = \frac{\sum_{i=1}^{n} g(Y_i) k\left[\frac{x - X_i}{\alpha_n}\right]}{\sum_{i=1}^{n} k\left[\frac{x - X_i}{\alpha_n}\right]} \qquad [8.3.12]$$

The extension of these estimators to the case where the observations form a dependent but stationary sequence has been accomplished by several authors (see, for example, Yakowitz (1985a,b), Collomb (1983, 1984), and Bosg (1983)). Suppose that Y_t is a time series process. Then [8.3.11] is an estimate of the marginal density function and if $X_i = Y_{i-1}$ then [8.3.12] is an estimate of $E[g(Y_t)|Y_{t-1} = y]$. The main condition for the use of the estimators [8.3.11] and [8.3.12] when Y_t is a stationary process is that they satisfy some kind of asymptotic independence such as geometric ergodicity (Yakowitz, 1985a). Note that if the process is Markov, $E[g(Y_t|Y_{t-1} = y)]$ is the optimal estimate of $g(Y_t)$ given the whole past under a least squares criterion. The main advantage of the estimators is the great flexibility that they provide to model nonlinearities when the nature of the departure from linearity is not obvious, as is the case in hydrological time series.

The higher order extensions of [8.3.11] and [8.3.12] are obvious and, hence, are not presented here. The choice of the parameter, α_n, is critical to obtain a balance between reduction of bias and reduction of variance of the estimates. The following procedure is employed to determine α_n for the models. For each point in the training set, one estimates the conditional regression function based on the rest of the training samples and obtains the sum of squares of the difference between the observed value and the estimate. This procedure is repeated for a range of values for α_n within which the absolute minimum of the sum of squares is found. The value of α_n which yields the minimum sum of squares is selected.

8.3.4 Forecasting Study

Introduction

To compare the forecasting performance of the various nonseasonal models mentioned in Section 8.3.3, two split sample experiments are performed. Annual river flow, tree ring indices, mud varve and annual temperature series are considered in these studies. Nonseasonal models are fitted to the first parts of the series and these models are then employed to forecast the remaining data.

Forecasting can, in fact, be used as a means of model discrimination among competing models. For a given type of data such as hydrological time series, select the class of models which forecast the best according to certain criteria. In economics, authors such as Granger and Newbold (1977) and Makridakis and Hibon (1979) have carried out extensive forecasting experiments to determine the best kinds of models to use with nonseasonal and seasonal data. Although water resources engineers have recognized the importance of forecasting for a long time, very few large forecasting studies have been executed. Consequently, the forecasting study presented in this section as well as by Noakes et al. (1988) and Noakes (1984) constitutes one of the first extensive forecasting studies in water resources. Forecasting experiments with seasonal and transfer function-noise models are given in Chapters 15 and 18, respectively.

A comprehensive approach for carrying out forecasting experiments is depicted in Figure 8.3.1. In the forecasting study reported here none of the series are first transformed before fitting the five models listed in Section 8.3.3 to the first part of the series. Furthermore, when forecasting the last part of the series, one step ahead forecasts are determined. As shown below both the ARMA and nonparametric regression model of Yakowitz forecast better than the other three kinds of models listed in the previous section. Finally, as demonstrated by the simulation experiments carried out in Section 10.6, ARMA models are capable of statistically preserving important historical statistics of annual geophysical time series.

First Forecasting Experiment

The annual data sets considered in the first study are listed in Table 8.3.1. The riverflow and temperature data are obtained from Yevjevich (1963) and Manley (1953), respectively. The most appropriate type of ARMA models to fit to the last two series in Table 8.3.1 are given in Table 5.4.1.

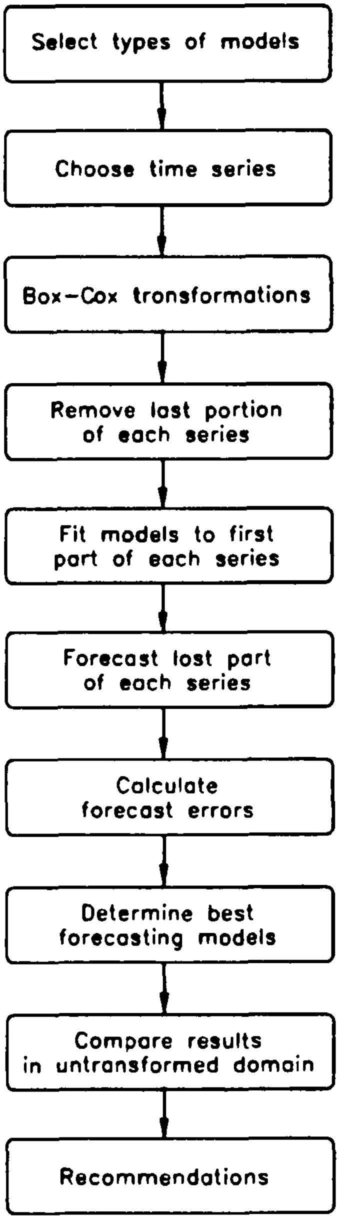

Select types of models
Choose time series
Box−Cox transformations
Remove last portion of each series
Fit models to first part of each series
Forecast last part of each series
Calculate forecast errors
Determine best forecasting models
Compare results in untransformed domain
Recommendations

Figure 8.3.1. Forecasting experiments.

Because of the computational effort required to forecast using the FGN and FDIFF models, only series with less than 150 data are considered in the first study. The general procedure is to truncate the data sets by omitting the last 30 years of data. Models are then calibrated using the first portion of the data. These models are then employed to forecast one step ahead MMSE forecasts (see Section 8.2) of the last 30 years of data. For a given model and time series, one can calculate the forecasting error for each of the 30 one step ahead forecasts. By summing the squared forecast errors, dividing by 30 and then taking the square root of this, one obtains the *root mean square error (RMSE)* for the forecasts.

The RMSE's for the 30 one step ahead MMSE forecasts for each of the models entertained are given in Table 8.3.2. A summary of these results is presented in Table 8.3.3. The rank sum is simply the sum of the product of the rank and the associated table entry. Thus, models with low rank sums forecast better overall than models with higher rank sums. In this study, the nonparametric model proposed by Yakowitz (1985a,b) forecasts well for the time series considered while the FDIFF model is the worst model.

Pitman's test (see Section 8.3.2) is employed to test for statistically significant differences in the RMSE's of the forecasts. The five competing procedures are compared in a pairwise fashion. The correlation values, r, are presented in Table 8.3.4. For these r values, the 95% confidence limits are calculated to be $\pm\frac{1.96}{\sqrt{30}} = \pm 0.358$. The ARMA, Markov, FGN and nonparametric forecasts are all significantly better at the 5% significance level of ± 0.358 than the FDIFF forecasts for the series Ogden. The nonparametric forecasts are also significantly (0.05 level) better than the FGN forecasts for the series Ogden.

Second Forecasting Experiment

The data sets employed in the second study are listed in Table 8.3.5. Except for the Snake time series, the tree ring indices are from Stokes et al. (1973). The Snake tree ring indices are from Schulman (1956). The most appropriate type of ARMA model to fit to the Navajo series is listed in Table 5.4.1 as being ARMA(1,1).

The RMSE's of the ARMA, Markov, nonparametric and FARMA forecasts are presented in Table 8.3.6 while a summary of these results is given in Table 8.3.7. In all cases, the Markov model has the largest RMSE of the four models considered in this study. The ARMA and nonparametric models are essentially equal in performance and are both slightly better than the FARMA model.

The likelihood ratio test described in Section 8.3.2 is employed to test for significant differences between the ARMA and Markov forecast errors. In this case, the test statistic, R, is calculated for both instances, where the means of the forecast errors are assumed to be zero ($R1$) and non-zero ($R2$). The calculated values are presented in Table 8.3.8. There is virtually no difference between $R1$ and $R2$ so either value may be employed in the test. In this study, the ARMA forecasts are significantly (0.05 level) better than the Markov forecasts for the two series Eaglecol and Lakeview. Since the RMSE's of the ARMA models are always less than the RMSE's of the Markov models, the Markov forecasts could never be significantly better than the ARMA forecasts.

Table 8.3.1. Annual riverflow and temperature data sets.

Code names	River or data types	Locations	Periods	n
Gota	Gota	Sjotorp-Vanersburg, Sweden	1807-1957	150
Mstouis	Mississippi	St. Louis, Missouri	1861-1957	96
Neumunas	Neumunas	Smalininkai, USSR	1811-1943	132
Ogden	St. Lawrence	Ogdensburg, New York	1860-1957	97
Temp	Temperature	Central England	1802-1951	150

Table 8.3.2. RMSE's for the one step ahead forecasts for the annual riverflow and temperature series.

Code names	ARMA	FGN	FDIFF	Markov	Nonparametric
Gota	87.58	95.57	97.66	97.45	92.86
Mstouis	1508.03	1543.56	1574.85	1625.90	1560.00
Neumunas	118.30	115.80	116.12	114.70	115.40
Ogden	473.89	630.55	875.91	450.85	426.90
Temp	1.21	1.17[a]	1.17	1.13	0.95

[a] Indicates smaller of tied values.

Table 8.3.3. Distribution of the RMSE's for 30 forecasts for the annual riverflow and temperature series.

Ranks	Number of times each model has indicated rank				
	ARMA	FGN	FDIFF	Markov	Nonparametric
1	2	0	0	1	2
2	0	1	0	2	2
3	1	3	0	0	1
4	0	1	3	1	0
5	2	0	2	1	0
Rank sum	15	15	22	14	9

Pitman's test is employed to compare the ARMA, nonparametric and FARMA forecasts in a pairwise fashion. The calculated correlations, r, between S_t and D_t are presented in Table 8.3.9. The only significant value (0.05 level) is for the series Lakeview when the ARMA and nonparametric forecasts are compared. Thus, the ARMA forecasts are significantly better than the nonparametric forecasts for this series at the 5% level. In all other cases, there is no statistically significant difference in the forecasts produced by the various models.

Table 8.3.4. Pitman's correlations, r, for pairwise comparisons of 5 annual models for each of the 5 series.

	Gota	Mstouis	Neumunas	Ogden	Temp
A vs B	-0.170	-0.112	0.125	-0.347	0.112
A vs C	-0.223	-0.171	0.089	-0.593	0.103
A vs D	-0.302	-0.317	0.102	0.076	0.165
A vs E	-0.277	-0.193	0.142	0.160	0.241
B vs C	-0.142	-0.275	-0.049	-0.828	-0.092
B vs D	-0.060	-0.209	0.040	0.335	0.142
B vs E	0.063	-0.123	0.041	0.453	0.209
C vs D	0.008	-0.114	0.053	0.582	0.143
C vs E	0.123	0.083	0.096	0.663	0.212
D vs E	0.178	0.167	-0.029	0.081	0.180

*Models: ARMA = A, FGN = B, FDIFF = C, Markov = D, Nonparametric = E.

Discussion

Based upon the result of the forecasting studies, the use of FGN and FDIFF models for forecasting annual hydrological and tree ring time series is not recommended. The two models which should be given serious consideration are the nonseasonal ARMA model and the nonparametric model presented by Yakowitz (1985a). Both forecast equally well for the series considered in the studies presented in this section. Moreover, Noakes (1989) demonstrates that the nonparametric model works well for generating inseason forecasts of salmon returns.

The performance of the various models is evaluated using the RMSE's of the forecasts and some of the statistical tests outlined in Section 8.3.2. This assumes that identical costs are assigned to both negative and positive forecast errors of the same magnitude. One recognizes that an asymmetric loss function may be more appropriate in certain instances, particularly in hydrological applications. For instance, different costs may be associated with inaccurate forecasts that result in either a flood or a drought. However, the RMSE criterion is employed since the procedures used for estimating the model parameters involve minimizing the sum of squared error terms. Presumably, if the type of loss function to be used to evaluate the forecast performance is known a priori, then the parameter estimation procedures could be adapted to minimize the expected loss. Without prior knowledge of the type of loss function, the RMSE criterion would appear to be a reasonable compromise (Noakes et al., 1985, 1988).

8.4 CONCLUSIONS

By following the model construction procedure of Part III, one can develop a parsimonious ARMA or other type of model for describing a given time series. As explained in Section 8.2, one can then use this model to produce MMSE forecasts of future observations. If one wishes to compare the forecasting accuracy of a range of models for a specified kind of time series, one can use the general model discrimination procedure outlined in Figure 8.3.1. By using tests from Section 8.3.2, one can ascertain if one model forecasts one step ahead forecasts significantly better than another. The results of the forecasting experiments of Section 8.3 demonstrate that ARMA models forecast annual hydrological and tree ring series just as well or better than any of

Table 8.3.5. Tree ring indices data.

Code names	Types of Trees	Locations	Periods	n
Bigcone	Bigcone spruce	Southern California	1458-1966	509
Dell	Limber pine	Dell, Montana	1311-1965	655
Eaglecol	Douglas fir	Eagle, Colorado	1107-1964	858
Exshaw	Douglas fir	Exshaw, Alberta	1460-1965	506
Lakeview	Ponderosa pine	Lakeview, Oregon	1421-1964	544
Naramata	Ponderosa pine	Naramata, B.C.	1415-1965	515
Navajo	Douglas fir	Navajo National Monument, Belatakin, Arizona	1263-1962	700
Ninemile	Douglas fir	Ninemile Canyon, Utah	1194-1964	771
Snake	Douglas fir	Snake River Basin	1282-1950	669

Table 8.3.6. RMSE's of the last half of the tree ring series forecasted.

Code names	ARMA	Markov	Nonparametric	FARMA
Bigcone	38.52	39.01	38.33	38.83
Dell	36.83	37.73	37.41	37.16
Eaglecol	27.73	29.00	28.11	27.60
Exshaw	32.70	33.58	32.51	32.77
Lakeview	16.75	17.78	17.11	16.86
Naramata	29.98	30.75	30.16	30.18
Navajo	44.27	44.46	44.17	44.39
Ninemile	38.18	38.53	37.93	37.78
Snake	21.87	22.43	21.74	21.78

Table 8.3.7. Distribution of the RMSE's for the ARMA, Markov, Nonparametric and FARMA models when the last half of the tree ring series forecasted.

Ranks	Number of times each model has indicated rank			
	ARMA	Markov	Nonparametric	FARMA
1	3	0	4	2
2	4	0	2	3
3	2	0	3	4
4	0	9	0	0
Rank sum	17	36	17	20

Table 8.3.8. ARMA vs Markov likelihood ratio statistics for the last half of the tree ring series forecasted.

Code names	$R1^a$	$R2^b$
Bigcone	0.587	0.587
Dell	2.160	2.157
Eaglecol	6.667	6.665
Exshaw	3.036	3.032
Lakeview	9.323	9.324
Naramata	2.056	2.053
Navajo	0.176	0.176
Ninemile	0.694	0.691
Snake	2.381	2.381

[a] The means of the forecast errors are assumed to be zero.

[b] The means of the forecast errors are not assumed to be zero.

Table 8.3.9. Pairwise comparison of the ARMA, Nonparametric and FARMA models using Pitman's test and forecasting the last half of the tree ring series.

Code names	ARMA vs Nonparametric	ARMA vs FARMA	FARMA vs Nonparametric
Bigcone	-3.79E-2	-6.49E-3	-7.52E-2
Dell	8.83E-2	2.92E-4	8.76E-2
Eaglecol	3.43E-2	4.37E-2	4.41E-2
Exshaw	-7.95E-2	-1.57E-2	-1.57E-2
Lakeview	1.21E-1[a]	-8.88E-3	5.61E-2
Naramata	7.94E-2	-1.09E-2	-7.25E-2
Navajo	-1.83E-2	-2.14E-3	-2.93E-2
Ninemile	-3.89E-2	1.83E-2	2.23E-2
Snake	-3.55E-2	5.36E-3	1.68E-3

[a] Significant at the 5% level.

its competitors. For this and many other reasons, ARMA models are highly recommended for use in practical applications.

When forecasting, it is important to use models that provide an adequate fit to the data using as few model parameters as possible. For certain types of models, Ledolter and Abraham (1981) demonstrate that if a nonparsimonious model is employed for forecasting, the variance of the forecast errors increases. This problem may not be serious for large samples but for a small number of observations the effect of overfitting may not be negligible.

When using the techniques of Section 8.2 to calculate MMSE forecasts, one assumes that the model parameters are known exactly. However, in practice one must estimate the model parameters from the data. The uncertainty contained in the parameter estimates could be considered when forecasting. Several authors (Akaike, 1970; Bloomfield, 1972; Bhansali, 1974; Box and Jenkins, 1976, p. 267; Baillie, 1979) present results for the variance of the forecast error when fitted parameters are used in various time series models under the unrealistic assumption that a forecasted data point is independent of the data employed for parameter estimation. Extending earlier work of Phillips (1979), Kheoh (1986) plus Kheoh and McLeod (1989) develop an expression for the l-step ahead forecast error of an AR(1) model for which the effect on the variance of the forecast error when the parameter is estimated from the same data upon which the forecast is based is taken into account. In particular, the effect of estimating the parameter is to cause a reduction in the variance of the forecast error.

Forecasting procedures similar to those developed for nonseasonal models, can also be extended for use with seasonal and other types of models. In Chapter 15, for example, MMSE forecasts are calculated and compared for three different types of seasonal models. Forecasting results for transfer-function noise models having one output series and multiple inputs, are presented in Chapter 18.

When one can select a range of models to fit to a time series, one may wish to select the model that forecasts most accurately. An alternative approach is to combine the forecasts from two or more models in accordance with their relative performances. In this way, one may be able to take advantage of the forecasting strengths of each of the models. In Section 15.5.2, specific techniques are developed for combining forecasts in an optimal manner from various models in order to attempt to improve the overall accuracy of the resulting forecasts. In addition, two case studies are presented in Sections 15.5.2 and 18.4.2 for examining the utility of combining forecasts. Similar combination techniques could, of course, be used with nonseasonal models.

PROBLEMS

8.1 For an ARMA(2,1) model, calculate MMSE forecasts up to a lead time of 10.

8.2 Determine MMSE forecasts for an ARIMA(1,2,1) model up to a lead time of 10.

8.3 As explained in Section 4.3.3, the most appropriate model to fit to the U.S. electrical demand series is an ARIMA(0,2,1) model with $\lambda = 0.5$. The last two data points in the transformed series are $z_t^{(\lambda)} = 2561$ and $z_{t-1} = 2491$ while $\hat{a}_t = -13.32$, $\sigma_a^2 = 636.7$ and $\hat{\theta}_1 = 0.9563$. Calculate by hand the MMSE forecasts along with their 50% probability limits up to five steps ahead from the last observation. Compare your results to Figure 8.2.3.

8.4 Suppose that an ARIMA(0,1,1) model is written as

$$\nabla z_t = a_t - 0.4a_{t-1}$$

Generate forecasts for lead times $l = 1,2,3$ from origin t when the model is written as given above, in random shock form and inverted form.

8.5 Suppose that the best model to fit to a time series represented by z_t, $t = 1,2, \ldots, n$ is

$$(1 - B)\ln z_t = (1 - \theta_1 B)a_t$$

From origin time t calculate the MMSE forecasts for $l = 1,2, \ldots, 6$ in both the transformed and untransformed domains.

8.6 Fit the most appropriate ARMA or ARIMA model to a series of your choice. Plot forecasts along with 50% probability limits from the last data point up to lead time $l = 20$.

8.7 Take a yearly series having at least 70 observations and fit an ARMA or ARIMA model to the reduced version of this series that omits the last 20 data points. Using the calibrated model, calculate one step ahead forecasts and plot them against the known observations. Using this graph and other appropriate calculations, comment upon how well your model forecasts.

8.8 On December 13, 1978, Makridakis and Hibon (1979) read their paper on an empirical investigation of forecasting accuracy before the Royal Statistical Society in London. Summarize some of the main empirical findings of these authors. At the end of Makridakis and Hibons' paper, comments by attendees are presented. Mention some of the more interesting criticisms made about their paper and comment upon how well the authors defended themselves.

8.9 Explain how parameter uncertainty can be considered when forecasting with ARMA models.

8.10 Why can a nonparsimonious model increase the variance of the forecast errors of MMSE forecasts generated by an ARMA model?

REFERENCES

DATA SETS

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society*, 79:242-261.

Schulman, E. (1956). *Dendroclimatic Changes in Semi-Arid America*. University of Arizona Press, Tucson, Arizona.

Stokes, M. A., Drew, L. G., and Stockton, C. W. (1973). Tree ring chronologies of Western America. Chronology Series 1, Laboratory of Tree Ring Research, University of Arizona, Tucson, Arizona.

United States Bureau of the Census (1976). The statistical history of the United States from colonial times to the present.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology paper no. 1, Colorado State University, Fort Collins, Colorado.

FORECASTING

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society, Series B,* 38(2):189-203.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series.* Academic Press, New York.

Kolmogorov, A. N. (1939). Sur l'interpolation et l'extrapolation des suites stationnaires. *C. R. Acad. Sci. Paris,* 208:2043-2045.

Kolmogorov, A. N. (1941a). Interpolation und extrapolation von stationaren zufalligen folgen. *Bull. Acad. Sci. (Nauk) U. R. S. S. Ser. Math.,* 5:3-14.

Kolmogorov, A. N. (1941b). Stationary sequences in Hilbert space. *(in Russian). Bull. Math. Univ. Moscow,* 2(6):1-40.

Ledolter, J. and Abraham, B. (1981). Parsimony and its importance in time series forecasting. *Technometrics,* 23(4):411-414.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting,* 1:111-153.

Makridakis, S. and Hibon, M. (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society, Series A,* 142:97-145.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society,* Series A, 137:131-165.

Whittle, P. (1963). Prediction and Regulation. *English Universities Press,* London.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* Wiley, New York.

Wold, H. O. (1954). *A Study in the Analysis of Stationary Time Series.* Almquist and Wicksell, Uppsala, Sweden, second edition.

HYDROLOGICAL FORECASTING

International Association of Hydrological Sciences (1980). Hydrological Forecasting Symposium, *Proceedings of the Oxford Symposium,* held April 15-18, 1980, IAHS-AISH Publication No. 129.

Noakes, D. J. (1984). Applied Time Series Modelling and Forecasting. PhD thesis, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Noakes, D. J. (1989). A nonparametric approach to generating inseason forecasts of salmon returns. *Canadian Journal of Fisheries and Aquatic Sciences,* 46(10):2046-2055.

Noakes, D. J., Hipel, K. W., McLeod, A. I., Jimenez, J. and Yakowitz, S. (1988). Forecasting annual geophysical time series. *International Journal of Forecasting,* 4:103-115.

Noakes, D. J., McLeod, A. I., and Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, 1:179-190.

NONPARAMETRIC MODELLING

Bosg, D. (1983). Nonparametric prediction in stationary processes. In *Lecture Notes in Statistics*, volume 16. Springer-Verlag, New York.

Collomb, G. (1983). From parametric regression to nonparametric prediction: Survey of mean square error and original results on the predictogram. In *Lecture Notes in Statistics*, volume 16. Springer-Verlag, New York.

Collomb, G. (1984). Proprietes de convergence presque complete du predictive a noyau. *Zeitschrift fur Wahrschein-lichkeitstheorie und Verwandte Gebiete*, 66:441-460.

Denny, J., Kisiel, C. and Yakowitz, S. J. (1974). Procedures for determining the order of Markov dependence in streamflow records. *Water Resources Research*, 10(5):947-954.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of density functions. *Annals of Mathematical Statistics*, 27:832-837.

Rosenblatt, M. (1971). Curve estimates. *Annals of Mathematical Statistics*, 42:1815-1842.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, Series A, 26:359-372.

Yakowitz, S. J. (1973). A stochastic model for daily river flows in arid regions. *Water Resources Research*, 9(5):1271-1285.

Yakowitz, S. J. (1976). Small-sample hypothesis tests of Markov order, with application to simulated hydrologic chains. *Journal of the American Statistical Association*, 71(353):132-136.

Yakowitz, S. J. (1979a). A nonparametric Markov model for daily river flow. *Water Resources Research*, 15(5):1035-1043.

Yakowitz, S. J. (1979b). Nonparametric estimation of Markov transition functions. *The Annals of Statistics*, 7:671-679.

Yakowitz, S. J. (1985a). Nonparametric density estimation and prediction for Markov sequences. *Journal of the American Statistical Association*, 80:215-221.

Yakowitz, S. J. (1985b). Markov flow models and the flood warning problem. *Water Resources Research*, 21(1):81-88.

PARAMETER UNCERTAINTY IN FORECASTING

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203-217.

Baillie, R. T. (1979). Asymptotic prediction mean squared error for vector autoregressive models. *Biometrika*, 66:675-678.

Bhansali, R. J. (1974). Asymptotic mean-square error of predicting more than one-step ahead using the regression model. *Applied Statistics*, 23:35-42.

Bloomfield, P. (1972). On the error prediction of a time series. *Biometrika*, 59:501-507.

Kheoh, T. S. (1986). Topics in Time Series Analysis and Forecasting. PhD thesis, Dept. of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada.

Kheoh, T. S. and McLeod, A. I. (1989). On the forecast error variance of fitted time series models. *Pakistan Journal of Statistics*, 5(3):239-243.

Phillips, P. C. B. (1979). The sampling distribution of forecasts from a first-order autoregression. *Journal of Econometrics*, 9:241-261.

STATISTICAL TESTING

Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburg, England.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31:9-12.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York, second edition.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. The Iowa State University Press, Ames, Iowa.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80-83.

CHAPTER 9

SIMULATING

WITH

NONSEASONAL MODELS

9.1 INTRODUCTION

Two of the most important uses of time series models in engineering are forecasting and simulation. In water resources applications, simulated sequences can assist designers in determining the appropriate size of a system and estimating the associated benefits and costs. Once in operation, reliable forecasts are required to ensure that the maximum benefits are obtained from operating the system. Chapter 8 of Part IV of the book deals with forecasting whereas the purpose of this chapter is to present flexible methods for simulating with ARMA and ARIMA models. Table 1.6.3 indicates the locations in the book where explanations are given about the theory and practice of both simulation and forecasting for different kinds of models.

The main objective of *forecasting* is to use the time series model fitted to a data set to obtain the most accurate estimate or prediction of future unknown observations. The goal of *simulation* is to employ the fitted model to generate a set of stochastically equivalent sequences of observations which could possibly occur in the future. These *simulated sequences* are often referred to as *synthetic data* by hydrologists because they are only possible realizations of what could take place. As a matter of fact, the overall science of fitting stochastic models to hydrologic data and using these models for simulation purposes is often called *synthetic hydrology.* Other titles for this field include *stochastic and operational hydrology.*

Simulation is now a widely accepted technique to aid in both the *design and operation* of water resources systems. Vogel and Stedinger (1988), for instance, demonstrate that using synthetic data generated by stochastic streamflow models can lead to improvements in the prediction of reservoir design capacity estimates. Besides the design and operation of large-scale engineering systems, another main use of simulation is to *investigate theoretical properties* of stochastic models. Often it is analytically impossible to derive certain theoretical characteristics of a given type of time series model. However, by using simulation one can determine these theoretical properties to any desired level of accuracy. In Chapter 10, simulation is used with ARMA models to study theoretical problems related to what is called the Hurst phenomenon.

When carrying out a simulation study, there are certain problems that should be avoided. For instance, many current simulation methods that are widely accepted, do not use correct initial values. Although the effect of *starting values* is transitory, it could cause systematic bias in a simulation study and, therefore, as pointed out by Moran (1959, Ch. 5) and Copa's (1966), the choice of initial values is important. To attempt to overcome this problem, some researchers discard the first section of a synthetic time series to supposedly get rid of the effects of initial values. However, exactly how many values of the generated series should be rejected and how much computer time is wasted by generating data that is not used?

As an example of a conservative approach to the effect of starting values, consider the simulation study of Brown and Hardin (1973). These authors used deterministic starting values for an AR model of order two and then generated a series with a length of 30,000 values. The first 15,000 values of the synthetic trace were discarded to supposedly nullify the effects of using non-random initial values.

The simulation procedures given in this chapter and also by McLeod and Hipel (1978) do not require fixed starting values. They are designed in a manner such that random realizations of the underlying stochastic process are used as initial values. Therefore, the results of a simulation study are not significantly biased and it is not necessary to disregard any of the generated data.

Often it is required to generate k' time series of length k. Some researchers resort to producing a single synthetic series of length $k' \cdot k$ and then splitting this long series into k' series of length k. If any serial correlation is present, then the results of any simulation study will be biased by this rather crude procedure. To overcome this problem of *bias*, the authors recommend generating k' separate time series of length k. If the generating procedures given in this chapter are adopted, then each time another series of length k is obtained, new random realizations of the stochastic process are used as starting values.

Figure 9.1.1 displays the *overall approach* for using a time series model for simulating data. If necessary, the given time series can be transformed using the Box-Cox transformation in [3.4.30] in order to alleviate problems with non-normality and/or non-constant variance discovered in the original data or the residuals of a model fitted to the untransformed time series. Following the three stages of model construction presented in Part III and summarized in Figure III.1, one can then develop an ARMA or ARIMA model for describing the transformed series or the original series when a Box-Cox transformation is not needed. This fitted model can then be used for generating synthetic sequences. When simulating data using a time series model, procedures described in Section 9.2 can be employed for generating the uncorrelated a_t terms in the ARMA or ARIMA model. The generated a_t's are used in either the WASIM1 or WASIM2 procedures of Sections 9.3 and 9.4, respectively, in order to simulate unbiased sequences from ARMA models that use proper starting values. If one is using an ARIMA model for simulation purposes, then the methods of Section 9.5 must be used to produce the nonstationary sequences. Finally, if the simulation model was fitted to the transformed series, then one must take the inverse Box-Cox transformation of the simulated values to obtain synthetic values that have the same units as the original data, as explained in Section 9.6.

When the approach of Figure 9.1.1 is employed for obtaining simulated sequences, it is assumed that the parameters of the model producing the synthetic data are known exactly. However, in practice one must estimate the model parameters and the uncertainty contained in the parameter estimates is reflected by their standard errors. The WASIM3 procedure of Section 9.7 can be used for incorporating *parameter uncertainty* into a simulation study. Following this, three practical applications are given in Section 9.8 to portray the effectiveness of the aforementioned simulation techniques.

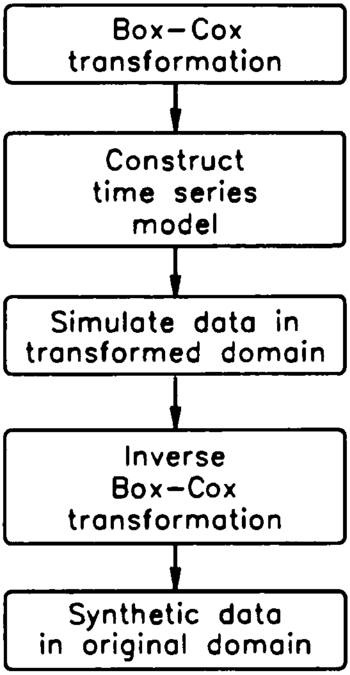

Box−Cox transformation
Construct time series model
Simulate data in transformed domain
Inverse Box−Cox transformation
Synthetic data in original domain

Figure 9.1.1. Overall procedure for simulating data.

9.2 GENERATING WHITE NOISE

9.2.1 Introduction

As shown in Figure 9.1.1, when developing a model for use in simulation, the first step is, if necessary, to transform the z_t series using the Box-Cox transformation in [3.4.30] to obtain the series $z_t^{(\lambda)}$. If the series is nonstationary, and, for example, the level is increasing over time, one can remove the nonstationarity using the differencing operator in [4.3.3] to obtain the stationary w_t series. As noted in Sections 2.4, 3.1, 4.1 and elsewhere in the book, often annual hydrological time series of moderate length are stationary. However, usually socio-economic series such as the water use, electricity consumption and Beveridge wheat price indices displayed in Figures 4.3.8, 4.3.10, and 4.3.15, respectively, are nonstationary. Whatever the case, one fits an ARMA model from [4.3.4] to the stationary w_t series which may be formed by first differencing the $z_t^{(\lambda)}$ series.

As an example as to how one would use an ARMA model for simulation, consider the cases of an AR(1) model written following [3.2.1] as

$$w_t = \phi_1 w_{t-1} + a_t \qquad [9.2.1]$$

where ϕ_1 is the first order AR parameter and a_t is the white noise sequence that is normal and independently distributed with mean zero and variance σ_a^2 (i.e. NID($0,\sigma_a^2$)). For convenience, it is assumed that the mean of w_t in [9.2.1] is zero. If the w_t series were formed by differencing the w_t series, the mean would be zero. However, if the $z_t^{(\lambda)}$ series were stationary and hence $z_t^{(\lambda)} = w_t$, then the mean would probably be nonzero. In either case, one can use [9.2.1] to simulate the w_t series. If there is a nonzero mean, it can simply be added to the simulated values.

Suppose one wishes to employ [9.2.1] to simulate 10 values of the w_t series. Then

$$w_2 = \phi_1 w_1 + a_2$$

Given that ϕ_1 is estimated from the data or else known in advance, one would have to have a starting value, w_1, and a normal white noise term, a_2, to simulate w_2. To avoid bias caused by using a fixed starting value for w_1, one can employ WASIM1 or WASIM2 from Section 9.3 and 9.4, respectively to obtain w_1. The white noise term a_2 is generated using an approach described in this section. Next, one can obtain w_3 using

$$w_3 = \phi_1 w_2 + a_3$$

where w_2 is known from the previous step and a_3 is generated by the computer. In general,

$$w_j = \phi_1 w_{j-1} + a_j\,, \quad j = 2,3, \ldots, 10$$

to simulate the ten values of w_t. If the w_t were differenced, one would have to use the algorithm of Section 9.5 to obtain the $z_t^{(\lambda)}$ series. Finally, the inverse Box-Cox transformation would have to be taken to get the simulated untransformed z_t's.

As shown in Figure 9.2.1, there are two main steps required to generate the a_t's which are NID($0,\sigma_a^2$). The first stage is to use an appropriate random number generator to produce independent random variables that follow a uniform distribution on the interval from zero to one. Random number generators are discussed in Section 9.2.2. The next step is to employ a technique that transforms the uniformly distributed variables to ones that follow the required distribution such as a normal distribution. The approach for accomplishing this is presented in Section 9.2.3. A classical text on generating independently distributed random variables is the one by Knuth (1969) while the book of Yakowitz (1977) provides a well written and entertaining account of this and other topics in computational probability and simulation.

9.2.2 Random Number Generators

Overview

As indicated in Figure 9.2.1, the first step in generating independently distributed random variables that follow the same distribution is to produce independent variables that follow a uniform distribution on the interval (0,1). To obtain uniformly distributed variables one uses what is called a *random number generator*. Following the instructions encoded into a computer program for the random number generator, a digital computer can produce the random numbers.

A digital computer follows a strictly deterministic process, since it exactly adheres to a program's precise instructions. Nonetheless, it is possible for a computer to generate a sequence of numbers $\{u_i\}$ which appear to be independent values distributed on the unit interval. These numbers are referred to as *pseudo-random numbers*. The *probability density function (pdf)*, $f(u)$, and the *cumulative distribution function*, $F(u)$, for a uniformly or rectangular distributed random variable, u, on the interval (0,1), are shown in Figure 9.2.2.

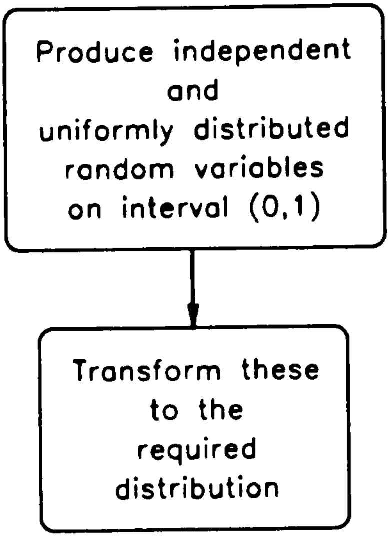
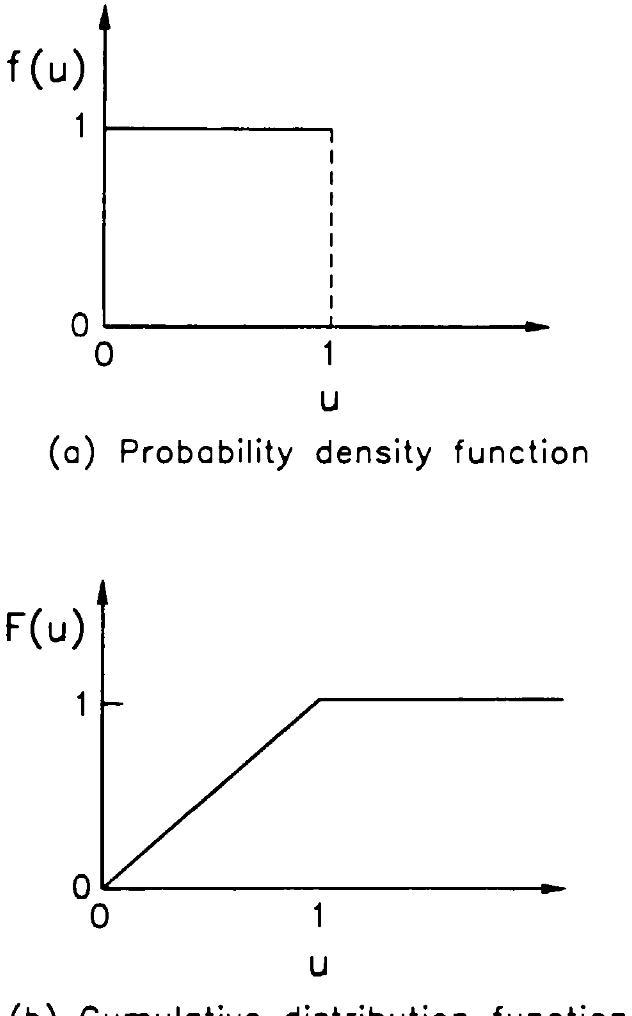

Produce independent and uniformly distributed random variables on interval (0,1)
Transform these to the required distribution

Figure 9.2.1. Generating identically and independently distributed random variables.

When employing a digital computer, some discrete approximation must be used in place of a continuous random variable such as one following the uniform distribution in Figure 9.2.2. Therefore, one may generate the first N terms of a decimal expansion of the value of a uniform variable. In an experiment for generating uniform variables, let D_j denote the jth decimal of the decimal expansion of an outcome. For each $k = 0,1,2,\ldots,9$, and $j = 1,2,\ldots,$ the event $D_j = k$ has a probability of $\frac{1}{10} = 0.1$.

The main idea behind a pseudo-random number generator is that successive numbers having the same length m are created in such a way that in the long run, each digit (i.e. $0,1,2,\ldots,9$) is expected to occur with probability 0.1 at each decimal place. Moreover, the occurrence of a given digit at a specific decimal place is independent of the digits occurring at other decimal places as well as previously generated random numbers.



(a) Probability density function

(b) Cumulative distribution function

Figure 9.2.2. Uniform probability density function and cumulative distribution function.

Over the years a range of pseudo-random number generators has been developed. One of the earliest random number generators is the *middle-square random number generator* of von Neumann (1951). One of the problems with this generator is that it has short cycles because after a certain length of time the uniform variables produced by the generator repeat themselves over relatively short time periods. Consequently, researchers developed generators that would have cycle lengths that are as long as possible.

Linear Congruential Random Number Generators

A particularly good family of pseudo-random number generator having maximum cycle length is the *linear congruential random number generator.* The linear congruential generators, originally suggested by Lehmer in 1948, form the most popular and highly studied class of methods for generating a sequence of pseudo-random numbers. These techniques are based upon the recurrence relationship

$$x_i = (ax_{i-1} + c)\ (mod\ m) \qquad [9.2.2]$$

where a *seed*, x_o, is an integer $0 \le x_o < m$ that is required to start the generator; multiplier a is an integer $0 < a < m$; increment c is an integer $0 \le c < m$. The symbol *mod m* stands for modulus m and, therefore, x_i is the remainder after $ax_{i-1} + c$ is divided by the integer m where $m > 0$. The sequence $\{x_i\}$ formed by allowing i to take on values $i = 1,2, \ldots,$ is often called a linear congruential sequence or a random number stream. In addition, when $c > 0$ or $c = 0$, the linear congruential generator is often referred to as a *mixed congruential generator* and a *multiplicative congruential generator*, (Wichmann and Hill, 1982), respectively.

The modulo arithmetic in [9.2.2] guarantees that each entry in $\{x_i\}$ is an integer falling in the interval $(0,m - 1)$. Consequently, the set of numbers, $\{x_i/m\}$, forms a sequence of uniformly distributed random variables, $\{u_i\}$.

One must carefully select the parameters a, c, m, and x_o in [9.2.2] for the linear congruential method in order to obtain a sequence that follows the distributional properties of independent uniformly distributed random variables, has a sequence containing the maximum period length, and achieves computational efficiency. The *cycle length or period* is some integer p such that $u_i = u_{i+p}$ for all $i \ge 0$. Clearly, this cycle length cannot be greater than the modulus m. A large period length and, hence, large m is required for achieving apparent randomness of the entries in the sequence $\{\mu_i\}$ or, equivalently, $\{x_i\}$. Hull and Dobell (1962) show that the linear congruential sequence $\{u_i\}$ has period m if and only if the following restrictions are satisfied:

1. c is relatively prime to m (i.e. c and m have no common factor other than unity).
2. $a - 1$ is a multiple of the period p, for every prime p dividing m.
3. $a - 1$ is a multiple of 4 if m is a multiple of 4.

Rules for finding the period for any choices of a, c and m are presented by Marsaglia (1972). Further detailed information regarding the selection of the parameter values in [9.2.2] is given by Janson (1966), Knuth (1969), Dieter (1972), as well as many other authors.

In summary, the algorithm for the linear congruential random number generator is as follows:

A. **Input:** Carefully select the parameters a, c and m in [9.2.2], plus the starting value x_o and the length N of the sequence to be generated.

B. **Calculations:**

0. Set $i = 1$
1. $x_i = (ax_{i-1} + c) \ (mod\ m)$
2. $u_i = truncated\ decimal\ expansion\ of\ x_i/m$ [9.2.3]
3. $i = i + 1$
4. If $i < N$, go to 1.
5. Stop.

C. **Output:** Sequence $\{u_i\} = u_1, u_2, \ldots, u_N$, of independent uniformly distributed random variables on the interval (0,1).

Example:

To demonstrate how the linear congruential random generator works, consider a simple illustration for which $a = 3$, $c = 7$, $m = 16$ and $x_o = 2$. Hence, equation [9.2.2] becomes

$$x_i = (3x_{i-1} + 7) \ (mod\ 16)$$

Using this equation, one calculates the sequence for x_i as:

$$x_o = 2 \ (starting\ value\ or\ seed)$$

$$x_1 = (3(2) + 7)(mod\ 16) = 13(mod\ 16) = 13$$

$$x_2 = (3(13) + 7)(mod\ 16) = 46(mod\ 16) = 14$$

$$x_3 = (3(14) + 7)(mod\ 16) = 49(mod\ 16) = 1$$

$$x_4 = (3(1) + 7)(mod\ 16) = 10(mod\ 16) = 10$$

$$x_5 = (3(10) + 7)(mod\ 16) = 37(mod\ 16) = 5$$

$$x_6 = (3(5) + 7)(mod\ 16) = 22(mod\ 16) = 6$$

$$x_7 = (3(6) + 7)(mod\ 16) = 25(mod\ 16) = 9$$

$$x_8 = (3(9) + 7)(mod\ 16) = 34(mod\ 16) = 2$$

$$x_9 = (3(2) + 7)(mod\ 16) = 13(mod\ 16) = 13$$

.
.
.

Consequently, the sequence $\{x_i\}$ is:

$$\{x_i\} = 2,13,14,1,10,5,6,9,2,13,\cdots$$

By dividing each entry in the x_i set of numbers by the modulus, one obtains the u_i sequence according to [9.2.3] as:

$$\{u_i\} = \frac{2}{16},\frac{13}{16},\frac{14}{16},\frac{1}{16},\frac{10}{16},\frac{5}{16},\frac{6}{16},\frac{9}{16},\frac{2}{16},\frac{13}{16},\cdots$$

Notice that these two sequences repeat themselves after only eight terms. Because of this, the entries in $\{u_i\}$ would not be independent. In addition, since the calculated values can only take on discrete values that are integer multiples of $\frac{1}{16}$ and many of these values such as $\frac{3}{16}$ and $\frac{4}{16}$ are missing, the u_i's are not uniformly nor continuously distributed. As noted earlier, to mimic independence and a uniform distribution, the period has to be as large as possible and, hence, a, c, m and x_o must be wisely selected.

Testing Random Number Generators:

As pointed out by Yakowitz (1977) and others, all feasible random number generators are inherently faulty from a purely philosophical viewpoint. Nonetheless, when appropriate choices of input parameters are made for the linear congruential random number generator, the generated sequences can satisfy standard statistical tests and are adequate for use in engineering applications. Traditionally, one employs separate tests to ascertain if the sequence $\{u_i\}$ is independent and also uniformly distributed. When using a random number generator on a given computer facility for the first time, one should invoke a range of statistical tests to ensure that the independence and uniform distribution assumptions of the u_i's are satisfied. Furthermore, one wishes to employ a random number generator that is also computationally efficient.

9.2.3 Generation of Independent Random Variables

General Approach

As shown in Figure 9.2.1, sequences following distributions other than the rectangular or uniform distribution are obtained by transforming rectangularly distributed random variables to the required distribution. For almost all of the models discussed in this book, one assumes that the a_t innovations are normally independently distributed as NID$(0,\sigma_a^2)$. Therefore, when simulating using an ARMA or ARIMA model, one wishes to transform the uniformly distributed variables generated using the techniques from the previous section into a sequence which is Gaussian or normally distributed and has uncorrelated elements.

More specifically, suppose one wishes to transform a sequence $\{u_i\}$ of independent uniformly distributed random variables to some other distribution such as a normal distribution. Let the variable following the other distribution be denoted by w where $f(w)$ and $F(w)$ represent the probability density function and cumulative distribution function, respectively. A ***universal random variable generator*** is available for transforming the uniform random variables to the required distribution. Following Yakowitz (1977, p. 41) the main steps in this algorithm are as follows:

A. **Inputs:** The sequence $\{u_i\}$ of independent uniformly distributed random variables and the formula $F(w)$ of the cumulative distribution for the distribution that one wishes the transformed variables to follow. Also, one should fix the length, N, of the w sequence to be generated.

B. **Calculations:**

```
0.  Set $i = 1$.
1.  For $u = u_i$, determine $w$ such that
        $w = minimum \left\{ y : F(y) \geq u \right\}$        [9.2.4]
2.  Assign $w_i = w$.
3.  $i = i + 1$
4.  If $i < N$, go to 1.
5.  Stop.
```

C. **Output:** Sequence $\left\{ w_i \right\} = w_1, w_2, \ldots, w_N$, which are independent and follow the required distribution $F(w)$. It can be proven that if the u_i's are uniform, the w_i's will follow the distribution $F(w)$.

Simulating Independent Normal Sequences

The probability density function for a *standard normal random variable* having a mean of zero and standard deviation of unity is defined by the probability density function

$$f(y) = (2\pi)^{-1/2}\exp(-y^2/2)\,,\quad -\infty < y < \infty \qquad [9.2.5]$$

In shortform notation, the distribution in [9.2.5] is written as $N(0,1)$. If the y's are also independently distributed, then they are normally independently distributed as NID(0,1). A normally distributed random variable, w, having a mean, μ_w, and standard derivation, σ_w, can be obtained from the standard normal variable, y, using the transformation

$$w = \sigma_w y + \mu_w \qquad [9.2.6]$$

The notation for the distribution of w is $N(\mu_w,\sigma_w^2)$. If the w's are normally independently distributed they are said to be NID(μ_w,σ_w^2). This *normal distribution* is often referred to as the *Gaussian distribution* in commemoration of the great German mathematician Karl Gauss.

A variety of approximate and exact algorithms are available for generating normally distributed random variables from uniformly distributed ones. Box and Muller (1958) provide an exact method for generating random variables. In particular, two random variables, u_1 and u_2, that are independent and uniformly distributed on the interval (0,1) (see Section 9.2.2), are transformed to $NID(0,1)$ random variables, y_1 and y_2, using the relationships

$$y_1 = (-2\ln u_1)^{1/2}\cos 2\pi u_2$$
$$y_2 = (-2\ln u_1)^{1/2}\sin 2\pi u_2 \qquad [9.2.7]$$

One then uses the transformation in [9.2.6] to obtain the random variables w_1 and w_2 that are distributed as $\mathrm{NID}(\mu_w,\sigma_w^2)$. The accuracy of the Box-Muller algorithm is determined by the computer word length as well as the accuracy of the logarithmic and trigonometric routines that are available on the computer. The calculations required in [9.2.7] can make this algorithm less computationally efficient relative to other competing algorithms.

A highly recommended algorithm for exactly generating NID(0,1) observations from independent uniformly distributed random variables is the method of Marsaglia and Bray (1964). Marsaglia et al. (1964) and also Knuth (1969, pp. 105-112) provide detailed information about how to encode the Marsaglia-Bray algorithm. Knuth (1969) thinks that the Marsaglia-Bray algorithm is more efficient from a computational viewpoint than the other perfect generator of Box and Muller (1958). However, the Marsaglia-Bray algorithm is more complicated and, therefore, requires a greater programming effort.

Generating Other Distributions

As pointed out in the first part of Section 9.2.3, a general approach is available for transforming random variables from a random number generator to any required distribution. In the previous subsection, for example, techniques are discussed for generating random variables that are NID(0,1). Methods are also available for generating independent variables which follow distributions such as exponential, gamma or Poisson distributions.

One can generate some types of independently distributed variables directly from the NID(0,1) variables. For instance, to generate independent *log-normal random variables*, the procedure is as follows:

1. Generate NID(0,1) random variables y_i, $i = 1,2,\ldots,N$.
2. Use the transformation

$$v_i = \exp(\mu_w + \sigma_w y_i) \qquad [9.2.8]$$

to obtain v_i, $i = 1,2,\ldots,N$, that are log normally distributed having parameters μ_w and σ_w for the mean and standard deviation, respectively, of the $\mathrm{NID}(\mu_w,\sigma_w^2)$ random variables.

Another example of generating a certain type of independent random variables directly from the NID(0,1) variables is the generation of independent *Pearson-type III variables*. The formula for the Pearson-type III distribution is

$$f(\mathrm{v}) = \frac{1}{\Gamma(p)}e^{-\mathrm{v}}\mathrm{v}^{p-1} \qquad [9.2.9]$$

where $0 \le \mathrm{v} < \infty, \ldots, p > 1$ and $\Gamma(p)$ is the gamma distribution. For generating independent *Pearson-type III random variables*, the procedure is as follows:

1. Generate NID(0,1) random variables y_i, $i = 1,2, \ldots, N$.

2. Use the transformation

$$v_i = p\left[1 - \frac{1}{9p} + \frac{y_i}{3\sqrt{p}}\right]^3 \qquad [9.2.10]$$

to obtain v_i, $i = 1,2, \ldots, N$, that are independent Pearson-type III distributed random variables.

Another approach for generating synthetic data is to use the empirical distribution of the given observations or some transformation thereof to simulate possible future occurrences. In Section 9.8.3, a case study is employed for explaining how the empirical distribution of the estimated innovations of a fitted ARMA model can be used to simulate the innovations.

9.3 WATERLOO SIMULATION PROCEDURE 1

Following the titles used by McLeod and Hipel (1978), the simulation procedures of Sections 9.3 and 9.4 are referred as *WASIM1 (Waterloo Simulation Procedure 1)* and WASIM2 (Waterloo Simulation Procedure 2), respectively. Both of these techniques avoid the introduction of bias into a simulated sequence by producing starting values that are randomly generated from the underlying stochastic process. WASIM1 consists of using the random shock or MA form of an ARMA model to simulate the starting values and then using the original ARMA model to simulate the remaining synthetic data.

Let w_t be a stationary w_t series for time $t = 1,2, \ldots, n$, to which an ARMA model is fitted as in [4.3.4] to produce the model

$$\phi(B)w_t = \theta(B)a_t \qquad [9.3.1]$$

where

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

is the nonseasonal autoregressive (AR) operator or polynomial of order p such that the roots of the characteristic equation $\phi(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the ϕ_i, $i = 1,2, \ldots, p$, are the nonseasonal AR parameters;

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

is the nonseasonal moving average (MA) operator or polynomial of order q such that the roots of $\theta(B) = 0$ lie outside the unit circle for invertibility and θ_i, $i = 1,2, \ldots, q$, are the nonseasonal MA parameters; the a_t's are identically independently distributed innovations with mean 0 and variance σ_a^2 [IID$(0,\sigma_a^2)$] and often the disturbances are assumed to be normally independently distributed [NID$(0,\sigma_a^2)$].

For simulation purposes, the zero-mean stationary seasonal ARMA model of Chapter 12 can be considered as a natural extension of the nonseasonal process. Models with a non-zero mean (or any other type of deterministic component) are simulated by first generating the corresponding zero-mean process and then adding on the mean component.

Suppose that the w_t's are expanded in terms of a pure MA process as in [3.4.18]. This is termed the random shock form of an ARMA process and is written as

$$w_t = \frac{\theta(B)}{\phi(B)}a_t = \psi(B)a_t = (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_t \qquad [9.3.2]$$

where $\psi_0 = 1$. A method for calculating the ψ_i parameters from the AR and MA parameters is presented in Section 3.4.3. If an AR operator is present, $\psi(B)$ forms an infinite series and therefore must be approximated by the finite series

$$\psi(B) \approx 1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_{q'} B^{q'} \qquad [9.3.3]$$

It is necessary to choose q' such that $\psi_{q'+1}, \psi_{q'+2}, \cdots$, are all negligible. Since the model is stationary, this can be accomplished by selecting q' sufficiently large such that the error given below is kept as small as desired.

$$\gamma_0 - \sum_{i=0}^{q'} \psi_i^2 < \mathit{error} \qquad [9.3.4]$$

where γ_o is the theoretical variance of a given ARMA process with $\sigma_a^2 = 1$ and is calculated using the algorithm of McLeod (1975) presented in Appendix A3.2; *error* is the chosen error level (ex. *error* $= 10^{-5}$).

To obtain a synthetic series of k observations, first generate $k + q'$ white noise terms $a_{-q'+1}, a_{-q'+2}, \ldots, a_0, a_1, a_2, \ldots, a_k$. Next, calculate

$$w_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots + \psi_{q'} a_{t-q'} \qquad [9.3.5]$$

where $t = 1,2,\ldots,r$, and $r = \max(p,q)$. The remaining w_t are easily determined from [9.3.1] as

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad [9.3.6]$$

where $t = r+1, r+2, \ldots, k$.

The use of [9.3.6] avoids the truncation error present in [9.3.3]. Nevertheless if an AR operator is present (i.e., $p > 0$), there will be some systematic error in the simulated data due to the approximation involved in [9.3.5]. However, this bias can be kept to a tolerable level by selecting the "error" term in [9.3.4] to have a specified minimum value. Of course, if the model is a pure MA(q) model, then set $q' = q$ and [9.3.5] is exact and can be utilized to generate all of the synthetic data.

An inherent advantage of the WASIM1 simulation technique is that the only restriction on the white noise terms is that they are IID($0,\sigma_a^2$). Although in many situations it is often appropriate to employ NID($0,\sigma_a^2$) innovations, this simulation method does not preclude considering other types of distributions. For instance, after modelling a relatively long hydrological time series, the residuals from the historical data could be used to form an empirical distribution function for generating the white noise. This approach is illustrated in the application in Section 9.8.3. In other situations, it may be warranted to simulate the white noise by employing *Johnson variates* (Johnson, 1949; Hill, 1976; Hill et al., 1976) which have been applied to hydrological data by authors including Sangal and Biswas (1970), Stedinger (1980) and Kottegoda (1987). Atkinson and Pearce (1976) discuss the computer generation of Beta, Gamma and normal random variables, while Delleur et al. (1976) suggest some distributions which can be employed in hydrology. Techniques for generating independent normal and other random variables are pointed out in Section 9.2.3.

9.4 WATERLOO SIMULATION PROCEDURE 2

9.4.1 WASIM2 Algorithm

WASIM2 (Waterloo Simulation Procedure 2) is based upon a knowledge of the theoretical autocovariance or autocorrelation function (ACF). Following McLeod (1975), a method for calculating the theoretical ACF for any ARMA process is presented in Section 3.4.2 and Appendix A3.2. Simulation approaches based upon knowing the theoretical ACF's of the underlying processes can also be used for simulating using other types of stochastic models. For example, in Section 10.4.6, the theoretical ACF of a fractional Gaussian noise model is used in the simulation technique for that model.

Suppose that it is required to generate k terms of an ARMA(p,q) model with innovations that are NID$(0,\sigma_a^2)$. The following simulation procedure is exact to simulate $w_1,w_2,\ldots,w_k$, for all stationary ARMA(p,q) models.

1. Obtain the theoretical autocovariance function γ_j for $j=0,1,\ldots,p-1$ by using the algorithm of McLeod (1975) with $\sigma_a^2=1$. (See Section 3.4.2 and Appendix A3.2.)

2. Following the approach of Section 3.4.3, determine the random shock coefficients ψ_j for $j=1,2,\ldots,(q-1)$ in [9.3.3].

3. Form the covariance matrix $\Delta\sigma_a^2$ of $w_p,w_{p-1},\ldots,w_1,a_p,a_{p-1},\ldots,a_{p-q+1}$.

$$\Delta=\begin{bmatrix}\left(\gamma_{i-j}\right)_{p\times p} & \left(\psi_{j-i}\right)_{p\times q}\\ \left(\psi_{i-j}\right)_{q\times p} & \left(\delta_{i,j}\right)_{q\times q}\end{bmatrix}_{(p+q)\times(p+q)} \qquad [9.4.1]$$

   In the above equation, the (i,j) element and dimension of each partitioned matrix are indicated. The values of $\delta_{i,j}$ are 1 or 0 according to whether $i=j$ or $i\neq j$, respectively. When $i-j<0$, then $\gamma_{i-j}=\gamma_{j-i}$ and $\psi_{i-j}=0$.

4. Determine the lower triangular matrix **M** by Cholesky decomposition (see Ralston (1965, p. 410), Healy (1968), or Hornbeck (1975)) such that

$$\Delta=\mathbf{M}\,\mathbf{M}' \qquad [9.4.2]$$

5. Generate $e_1,e_2,\ldots,e_{p+q}$, and $a_{p+1},a_{p+2},\ldots,a_k$ where the e_t and a_t sequences are NID$(0,\sigma_a^2)$.

6. Calculate $w_1,w_2,\ldots,w_p$, from

$$w_{p+1-t}=\sum_{j=1}^{t}m_{t,j}e_j\,,\quad t=1,2,\ldots,p \qquad [9.4.3]$$

   where $m_{t,j}$ is the t,j entry in the matrix **M**.

7. Determine $a_{p-q+1}, a_{p-q+2}, \ldots, a_p$, from

$$a_{p+1-t} = \sum_{j=1}^{p+t} m_{t+p,j} e_j \, , \quad t = 1,2, \ldots, q \qquad [9.4.4]$$

8. Obtain $w_{p+1}, w_{p+2}, \ldots, w_k$, using

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$
$$t = p+1, p+2, \ldots, k \qquad [9.4.5]$$

9. If another series of length k is required then return to step 5.

For a particular ARMA model, it is only necessary to calculate the matrix **M** once, no matter how many simulated series are synthesized. Therefore, WASIM2 is economical with respect to computer time required, especially when many time series of the same length are generated.

Often the white noise disturbances can be assumed to be NID$(0,\sigma_a^2)$ and it is desirable to have as much accuracy as possible in order to eliminate bias. For this situation, the authors recommend using WASIM2 for a pure AR model or an ARMA process. When simulating a pure MA process with innovations that are NID$(0,\sigma_a^2)$, the WASIM1 and WASIM2 procedures are identical.

9.4.2 Theoretical Basis of WASIM2

In Step 3 of the WASIM2 algorithm, one forms the covariance matrix $\Delta\sigma_a^2$ of $w_p, w_{p-1}, \ldots, w_1, a_p, a_{p-1}, \ldots, a_{p-q+1}$, which are contained in a vector W. Next, one determines the lower triangular matrix M for Δ in [9.4.2]. Following Steps 6 and 7, the starting values contained in the vector W can be generated using

$$\mathrm{W} = \mathrm{M\,e} \qquad [9.4.6]$$

where the e_t's contained in the vector e are NID$(0,\sigma_a^2)$. In order to simulate exactly the starting values contained in W, the covariance matrix of W must be $\Delta\sigma_a^2$. This can be easily proven as follows:

$$\begin{aligned} Var[\mathrm{M\,e}] &= E[\mathrm{M\,e}(\mathrm{M\,e})^T] \\ &= E[\mathrm{M\,e\,e}^T\mathrm{M}^T] = \mathrm{M}Var(\mathrm{e})\mathrm{M}^T \\ &= \sigma_a^2 \mathrm{M\,M}^T = \Delta\sigma_a^2 \end{aligned} \qquad [9.4.7]$$

where σ_a^2 is a diagonal matrix for which each diagonal entry is σ_a^2.

9.4.3 ARMA(1,1) Simulation Example

The purpose of this section is to show in detail for a specific model how calculations are made in the WASIM2 algorithm. Suppose one wishes to simulate 10 values with an ARMA(1,1) model using WASIM2. For convenience, assume that the mean of the w_t series in [9.3.1] is zero. If the mean were nonzero, it could be added to each of the simulated values.

The calculations for an ARMA(1,1) model using WASIM2 are as follows:

1. For an ARMA(1,1) model $p = 1$ and, hence, one only has to calculate γ_0 at step 1. From [3.4.16]

$$\gamma_0 = \frac{1+\theta_1^2-2\phi_1\theta_1}{1-\phi_1^2}\sigma_a^2$$

Letting $\sigma_a^2 = 1$

$$\gamma_0 = \frac{1+\theta_1^2-2\phi_1\theta_2}{1-\phi_1^2}$$

2. Using the identity in [3.4.21] one can calculate the random shock coefficients. Only $\psi_0 = 1$ is required for an ARMA(1,1) model.

3. The matrix Δ in [9.4.1] is

$$\Delta = \begin{bmatrix} \gamma_0 & \psi_0 \\ \psi_0 & \delta_{0,0} \end{bmatrix} = \begin{bmatrix} \gamma_0 & 1 \\ 1 & 1 \end{bmatrix}$$

Then $\Delta\sigma_a^2$ is the covariance matrix for (w_1,a_1).

4. In this step, one obtains the Cholesky decomposition matrix $\mathbf{M}$ for Δ such that

$$\Delta = \mathbf{M}\,\mathbf{M}^T$$

For the case of an ARMA(1,1) model

$$\Delta = \begin{bmatrix} \gamma_0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} m_{11} & 0 \\ m_{21} & m_{22} \end{bmatrix}\begin{bmatrix} m_{11} & m_{21} \\ 0 & m_{22} \end{bmatrix}$$

or

$$\begin{bmatrix} \gamma_0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} m_{11}^2 & m_{11}m_{21} \\ m_{21}m_{11} & m_{21}^2+m_{22}^2 \end{bmatrix}$$

By equating (i,j) entries in the matrices on the left and right hand sides of the above equations, one can calculate m_{11}, m_{21} and m_{22}. In particular, for the (1,1) element:

$$m_{11}^2 = \gamma_0$$

Therefore,

$$m_{11} = \sqrt{\gamma_0}$$

For the (1,2) entry:

$$m_{11}m_{21} = 1$$

Therefore,

$$m_{21} = \frac{1}{m_{11}} = \frac{1}{\sqrt{\gamma_0}}$$

for the (2,2) element:

$$m_{21}^2 + m_{22}^2 = 1$$

Therefore,

$$m_{22}^2 = 1 - m_{21}^2 = 1 - \frac{1}{\gamma_0}$$

Therefore,

$$m_{22} = \sqrt{1 - \frac{1}{\gamma_0}}$$

Hence,

$$\mathbf{M} = \begin{bmatrix} \sqrt{\gamma_0} & 0 \\ \frac{1}{\sqrt{\gamma_0}} & \sqrt{1 - \frac{1}{\gamma_0}} \end{bmatrix}$$

5. The techniques of Section 9.2.3 can be used to generate (e_1,e_2) and $(a_2,a_3,\ldots,a_{10})$ where the sequences are NID$(0,\sigma_a^2)$.

6. The starting value w_1 is calculated using [9.4.3] as

$$w_1 = m_{11}e_1 = \sqrt{\gamma_0}e_1$$

7. The initial value for a_1 is found using [9.4.4] to be

$$a_1 = m_{21}e_1 + m_{22}e_2 = \frac{1}{\sqrt{\gamma_0}}e_1 + \left(\sqrt{1 - \frac{1}{\gamma_0}}\right)e_2$$

8. Use the given definition of the ARMA(1,1) model in [9.4.5] to get $w_2,w_3,\ldots,w_{10}$. More specifically,

$$w_2 = \phi_1 w_1 + a_2 - \theta_1 a_1$$

where the starting values for w_1 and a_1 are calculated in steps 6 and 7, respectively. The values for a_2 and also $a_3, a_4, \ldots, a_{10}$ are determined in step 5. Next

$$w_3 = \phi_1 w_2 + a_3 - \theta_1 a_2$$

where w_2 is calculated in the previous iteration and a_3 and a_2 are generated at step 5. By following the same procedure

$$w_4 = \phi_1 w_3 + a_4 - \theta_1 a_3$$

$$\begin{matrix} \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{matrix}$$

$$w_{10} = \phi_1 w_9 + a_{10} - \theta_1 a_9$$

If w_t has a mean level, this can be added to the above generated values. Notice that the starting values w_1 and a_1 are randomly generated from the underlying ARMA(1,1) process. Therefore, the simulated sequence is not biased because of fixed starting values.

9.5 SIMULATION OF INTEGRATED MODELS

9.5.1 Introduction

As discussed in 2.4 and elsewhere in the book, for annual geophysical time series of a moderate length (perhaps a few hundred years), it is often reasonable to assume that a stationary model can adequately model the data. In Section 10.6 and also in Hipel and McLeod (1978), for example, stationary ARMA models are fitted to 23 time series which are measured from six different natural phenomena. Nevertheless, certain types of time series that are used in water resources could be nonstationary. The average annual cost of hydro-electric power and the total annual usage of water-related recreational facilities constitute two types of measurable processes which possess mean levels and variances that could change significantly over time. Other examples of nonstationary annual series are given in Section 4.3.3. In general, time series that reflect the socio-economic aspects of water resources planning may often be nonstationary, even over a short time span. Consequently, in certain situations it may be appropriate to incorporate a nonseasonal differencing operator into the nonseasonal model in order to account for the nonstationarity by following the approach of Section 4.3.

As explained in Chapter 12, if a seasonal ARIMA or SARIMA model is fit to seasonal data, usually both nonseasonal and seasonal differencing are required to account for the nonstationarity. Consider the case of average monthly observations. If the monthly mean and perhaps variance change from one year to the next for each specific month, then fitting a nonstationary SARIMA model to the data may prove to be reasonable. For example, the average monthly water demand for large cities tends to increase from year to year for each month. For the aforementioned situations, the simulation procedures for integrated models presented in this section could be useful.

As pointed out in Part VI, when considering seasonal hydrological data, such as average monthly riverflows, the individual monthly averages may have constant mean values but the means vary from month to month. Consequently, the time series of all the given data is by definition nonstationary but it still may not be appropriate to employ a nonstationary SARIMA model to describe the data. Rather, the given natural time series is firstly deseasonalized to produce a stationary nonseasonal data set and subsequently a nonseasonal model is fit to the deseasonalized data. For example, prior to fitting a nonseasonal ARMA model to the data, it is a common procedure to standardize average monthly riverflow time series to eliminate seasonality (see Chapter 13). The WASIM1 and WASIM2 simulation procedures of this chapter can be used with the deseasonalized models of Chapter 13 and the periodic models of Chapter 14. However, for both of these types of seasonal models, one does not have to difference the data and, hence, the methods of Section 9.5 do not have to be used when using them for simulation.

Nonseasonal ARIMA and SARIMA models are presented in Chapters 4 and 12, respectively. For completeness of presentation of this simulation chapter, simulating with both types of models is discussed below.

9.5.2 Algorithms for Nonseasonal and Seasonal ARIMA Models

Although caution should be exercised when modelling nonstationary data, it is evident that situations may arise when it is suitable to invoke *differencing*. Any seasonal ARIMA model from [12.2.7] that contains a differencing operator is termed an *integrated* model. Suppose that it is required to simulate k values of $z_t^{(\lambda)}$ by using an integrated process. A stationary w_t series is related to the nonstationary $z_t^{(\lambda)}$ series by the equation

$$w_t = \nabla^d \nabla_s^D z_t^{(\lambda)}, \quad t = d'+1, d'+2, \ldots, k \qquad [9.5.1]$$

where s is the seasonal length ($s = 12$ for monthly data); $\nabla^d = (1 - B)^d$ is the nonseasonal differencing operator of order d in [4.3.3] to produce nonseasonal stationarity of the dth differences and usually $d = 0, 1$ *or* 2; $\nabla_s^D = (1 - B^s)^D$ is the seasonal differencing operator of order D in [12.2.3] to produce seasonal stationarity of the Dth differenced data and usually $d = 0$, 1, or 2, and for nonseasonal data $D = 0$; $d' = d + sD$.

Because of the differencing in [9.5.1], the d' initial values $w_1, w_2, \ldots, w_{d'}$, which determine the "current level" of the process, are assumed known. Given the d' initial values, the time series integration algorithm forms the integrated series $z_t^{(\lambda)}$ for $t = d'+1, d'+2, \ldots, k$. The integrated series is derived theoretically from the relationship

$$z_t^{(\lambda)} = S^d S_s^D w_t \qquad [9.5.2]$$

where $S = \nabla^{-1} = 1 + B + B^2 + \cdots$, is the nonseasonal summation operator; $S_s = \nabla_s^{-1} = 1 + B^s + B^{2s} + \cdots$, is the seasonal summation operator.

When employing [9.5.2] to obtain an integrated series, the methods of the previous sections are utilized to determine the w_t sequence. Then the integration algorithm that is developed presently in this section, is used to evaluate [9.5.2]. The situation where it is required to simulate data from a nonseasonal model containing a differencing operator, is first considered. This is followed by a discussion of the generation of synthetic data from a general seasonal model that possesses a seasonal differencing operator and perhaps also a nonseasonal differencing operator.

Nonseasonal Model:

The integration algorithm for a nonseasonal ARIMA model (i.e. $s = D = 0$) is as follows:

For $i = 1, 2, \ldots, d$:

1. Determine the starting value $\nabla^{d-i} z_t^{(\lambda)}$ by differencing the given initial values $z_1^{(\lambda)}, z_2^{(\lambda)}, \ldots, z_d^{(\lambda)}$.

2. Calculate $\nabla^{d-i} z_t^{(\lambda)}$ for $t = d+1, d+2, \ldots, k$, by employing the identity

$$\nabla^{d-i} z_t^{(\lambda)} = \nabla^{d+1-i} z_t^{(\lambda)} + \nabla^{d-i} z_{t-1}^{(\lambda)} \qquad [9.5.3]$$

Seasonal Model:

For a seasonal model, the integration algorithm is subdivided into two parts. The first stage consists of performing the nonseasonal integration.

For $i = 1, 2, \ldots, d$:

1. Determine the starting value $\nabla^{d-i} \nabla_s^D z_d{}'$, by differencing the given initial values $z_1^{(\lambda)}, z_2^{(\lambda)}, \ldots, z_{d'}^{(\lambda)}$.
2. Calculate $\nabla^{d-i} \nabla_s^D z_t^{(\lambda)}$ for $t = d'+1, d'+2, \ldots, k$, by using the equation

$$\nabla^{d-i} \nabla_s^D z_t^{(\lambda)} = \nabla^{d+i-i} \nabla_s^D z_t^{(\lambda)} - \nabla^d - i \nabla_s^D z_{t-1}^{(\lambda)} \qquad [9.5.4]$$

In the second stage the seasonal integration is performed. For $i = 1, 2, \ldots, D$:

1. Determine the starting values $\nabla_s^D z_t^{(\lambda)}$ for $t = d', d'-1, \ldots, d' = s$, by differencing the given initial values $z_1^{(\lambda)}, z_2^{(\lambda)}, \ldots, z_{d'}^{(\lambda)}$.
2. Calculate $\nabla_s^{D-i} z_t^{(\lambda)}$ for $t = d'+1, d'+2, \ldots, k$ by using the equation

$$\nabla_s^{D-i} z_t^{(\lambda)} = \nabla_s^{D-i+1} z_t^{(\lambda)} + \nabla_s^{D-i} z_{t-s}^{(\lambda)} \qquad [9.5.5]$$

9.6 INVERSE BOX-COX TRANSFORMATION

As shown in Figure 9.1.1, before fitting a time series model to a given time series, one may wish to transform the data using the Box-Cox transformation in [3.4.30]. The main purposes of the Box-Cox transformation are to make the residuals of the fitted ARMA or ARIMA model (see Section 3.4.5) to be normally distributed and homoscedastic (have constant variance). Subsequent to simulating data sequences using the fitted model and following the techniques from the previous sections of this chapter, one must take the inverse Box-Cox transformation in order to obtain synthetic data that have the same units as the original time series.

From [3.4.30] the Box-Cox transformation of the original z_t series is

$$z_t^{(\lambda)} = \begin{cases} \lambda^{-1}[(z_t + c)^\lambda - 1], & \lambda \neq 0 \\ \ln(z_t + c) & , \lambda = 0 \end{cases} \qquad [9.6.1]$$

where the constant c is chosen to be just large enough to cause all of the entries of the z_t series to be positive. When the data are nonstationary, one may also wish to difference the $z_t^{(\lambda)}$ series to obtain the w_t series in [4.3.3]. Of course, if the z_t or $z_t^{(\lambda)}$ series are already stationary, the w_t series is the same as $z_t^{(\lambda)}$. Whatever the case, after following the methods of the previous sections of Chapter 9 to obtain the simulated $z_t^{(\lambda)}$ sequences, one must take the inverse Box-Cox transformation to get the synthetic z_t data. The inverse Box-Cox transformation is written as

$$z_t = \begin{cases} [\lambda z_t^{(\lambda)} + 1]^{1/\lambda} - c \, , & \lambda \neq 0 \\ \exp[z_t^{(\lambda)}] - c & , \lambda = 0 \end{cases} \qquad [9.6.2]$$

where $z_t^{(\lambda)}$ stands for the simulated sequence obtained directly from the ARMA or ARIMA model. If the $z_t^{(\lambda)}$ sequence possesses a mean level, one can add the transformed mean to each of the generated z_t values calculated using [9.6.2].

9.7 WATERLOO SIMULATION PROCEDURE 3

9.7.1 Introduction

In simulation studies, one usually employs a calibrated ARMA model to simulate possible future sequences of the time series to which the model is fitted. One could, of course, assume a given theoretical model having specified parameter values and then use this model in a simulation study. However, when fitting the model to the series, one obtains maximum likelihood estimates (MLE's) and standard errors (SE's) for the model parameters (see Section 6.2 and Appendices A6.1 and A6.2). Because the time series used to calibrate the ARMA model is only one finite realization of the underlying stochastic process that generated this observed sequence, the population values of the model are not known. The MLE's of the model parameters constitute the best estimates of the population values given the available information. The uncertainties or variations of these estimates are reflected by their SE's. As explained in Appendix A6.2, the SE's are calculated as the square roots of the diagonal entries in the variance-covariance matrix for the estimated parameters.

The *WASIM3 (Waterloo Simulation Procedure 3) algorithm* can be used in simulation studies where it is required to incorporate *parameter uncertainty* into the analysis. Suppose that it is necessary to generate k' synthetic traces of length k. When generating each series of length k, different values of the model parameters are randomly selected if WASIM3 is employed. The WASIM3 procedure is explained in this book only for a nonseasonal ARMA model, since extension to the seasonal case is straightforward. In terms of Figure 9.1.1, the WASIM3 algorithm is used prior to taking the inverse Box-Cox transformation discussed in Section 9.6. Following the presentation of the WASIM3 algorithm in the next subsection, it is explained how parameter uncertainty is incorporated into reservoir design. Section 9.7.4 discusses how practitioners can deal with model uncertainty.

9.7.2 WASIM3 Algorithm

Suppose that the historical time series containing N values is modelled as an ARMA(p,q) model as in [9.3.1] that has an estimated mean level of $\hat{\mu}$. The Gaussian white noise residuals have an estimated variance denoted by $\hat{\sigma}_a^2$. Let the vector of the estimated ARMA parameters be given by

$$\hat{\beta} = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q) \qquad [9.7.1]$$

(see Section 6.2). The vector of the true model parameters is denoted by

$$\beta = (\phi_1,\phi_2,\ldots,\phi_p,\theta_1,\theta_2,\ldots,\theta_q) \qquad [9.7.2]$$

The mean level of the true model is μ while the variance of the white noise is σ_a^2.

If a non-informative prior distribution is used for the model parameters, then β, μ and σ_a^2 are approximately independent with posterior distributions given by

$$\beta \approx N(\hat{\beta},\mathbf{V}_{\hat{\beta}}) \qquad [9.7.3]$$

where $\mathbf{V}_{\hat{\beta}}$ is the estimated variance covariance matrix of β which is usually calculated at the estimation stage of model development (see Appendix A6.2) and N means normally distributed.

$$\mu \approx N\left(\hat{\mu}, \left[\frac{1-\hat{\phi}_1-\hat{\phi}_2-\cdots-\hat{\phi}_p}{1-\hat{\theta}_1-\hat{\theta}_2-\cdots-\hat{\theta}_q}\right]^{-2} , \frac{\hat{\sigma}_a^2}{N}\right) \qquad [9.7.4]$$

$$\sigma_a^2 \approx N\left(\hat{\sigma}_a^2, \frac{\hat{\sigma}_a^4}{2}\right) \qquad [9.7.5]$$

The findings given in [9.7.3] to [9.7.5] are based upon large sample theory. Nevertheless, these results can be used to obtain some idea of the importance, if any, of parameter uncertainty in a particular situation. It should be noted that if an informative prior distribution were used, the variances of the parameters would be less and hence the parameter uncertainty would also decrease.

The following algorithm for WASIM3 can be used to allow for parameter uncertainty when k' series of length k are to be generated from an ARMA(p,q) model.

1. Set $i = 1$.

2. Randomly generate values for β, μ and σ_a^2 using the posterior distributions given in [9.7.3], [9.7.4] and [9.7.5], respectively. Denote the generated parameter values as β_i, μ_i and $\sigma_{a,i}^2$. Refer to the book by Janson (1966) for a method to obtain random values from a multivariate normal distribution.

3. Use WASIM2 (or WASIM1) for an ARMA(p,q) process with parameters β_i, μ_i, and $\sigma_{a,i}^2$, to simulate a synthetic series of length k that is represented by $z_1^{(i)},z_2^{(i)},\ldots,z_k^{(i)}$. If the model contains a Box-Cox transformation, the inverse transformation in [9.6.2] is required.

4. Set $i = i + 1$. If $i \leq k'$ then repeat steps 2 and 3 to obtain another possible realization of the time series. When $i > k'$, the WASIM3 procedure is terminated.

9.7.3 Parameter Uncertainty in Reservoir Design

In this section, an algorithm is presented for estimating the expected utility of a ***reservoir design*** given the specified ARMA(p,q) model for the riverflow data and a posterior distribution $p(\beta,\mu,\sigma_a^2)$ for the parameters. For a given riverflow time series $z_1,z_2,\ldots,z_k$, and a particular reservoir design D, the (vector-valued) net benefit function is given by

$$NB = NB(z_1, \ldots, z_k; D) \qquad [9.7.6]$$

and the utility is

$$U = U(NB) \qquad [9.7.7]$$

The expected utility of D is then given by

$$u(D) = E\left\{U(NB(z_1, z_2, \ldots, z_k; D))\right\}$$

$$= \int_{z_1}\int_{z_2} \cdots \int_{z_k}\int_{\beta_1}\int_{\beta_2} \cdots \int_{\beta_{p+q}}\int_{\mu}\int_{\sigma_a^2} U(NB(z_1, z_2, \ldots, z_k; D))$$

$$p(z_1, z_2, \ldots, z_k | \beta, \mu, \sigma_a^2) p(\beta, \mu, \sigma_a^2)$$

$$dz_1 dz_2 \cdots dz_k d\beta_1 d\beta_2 \cdots d\beta_{p+q} d\mu d\sigma_a^2 \qquad [9.7.8]$$

The best design, D_o, maximizes the value of $u(D)$.

After an ARMA model is fit to the given time series of historical riverflows, the following algorithm may be used to estimate $u(D)$ and a confidence interval (or Bayesian probability interval) for $u(D)$.

1. Set $i = 1$, $T_1 = 0$, $T_2 = 0$. Let k' be the number of series of length k that are to be generated. For example, k' may have a value of 10,000.
2. Generate a synthetic time series, $z_1^{(i)}, z_2^{(i)}, \ldots, z_k^{(i)}$, using the WASIM3 algorithm.
3. Calculate $u_i = U(NB(z_1^{(i)}, z_2^{(i)}, \ldots, z_k^{(i)}; D))$, set $T_1 = T_1 + u_i$, and set $T_2 = T_2 + u_i^2$.
4. Set $i = i + 1$ and go to step 2 if $i \le k'$. Go to step 5 if $i > k'$.
5. Set

$$\bar{u} = \frac{1}{k'} T_1 \qquad [9.7.9]$$

and let

$$S_{\bar{u}} = \left[\left(\frac{1}{k'} T_2 - \bar{u}^2\right) / k'\right]^{1/2} \qquad [9.7.10]$$

The calculated $\bar{u}$ provides an estimate of $u(D)$ and a 95% confidence interval (or Bayesian probability interval) for $u(D)$ is given by $\bar{u} \pm 1.96 S_{\bar{u}}$. Although the aforesaid algorithm is explained for a nonseasonal ARMA(p,q) model, the same approach is valid for seasonal models. The number of generated synthetic traces (i.e. k') can be increased if more accuracy is required or decreased when less accuracy is needed.

9.7.4 Model Uncertainty

In the synthetic hydrology approach to reservoir design and operation, an ARMA model may be fit to a historical riverflow time series and then used to simulate other possible realizations of the riverflows. Two sources of possible error may arise. The model selected may be inappropriate or the estimated parameters may be inaccurate. The procedures of Part III emphasize techniques for selecting an appropriate model followed by efficient parameter estimation and diagnostic checking for possible model inadequacies. It is thus reasonable to suppose that the selected model is at least approximately valid and that the best possible estimates are obtained for the model parameters by using the method of maximum likelihood (see Section 6.2). On the other hand, if a possible inappropriate model is fit to the data and no checks of model adequacy are done, a seriously inadequate model may be chosen. It is demonstrated in Section 10.4.6, for example, that the use of fractional Gaussian noise models may give poor fits to annual riverflow time series when compared to ARMA models. If the methods of Part III are used with a hydrologic time series of at least 50 observations, the selection of an inappropriate model is not likely to occur. The reader may wish to refer to Section 5.2 for further discussions of modelling philosophies and different kinds of uncertainties.

Given that the best possible model is identified for fitting to a series and efficient parameter estimates are obtained, the WASIM3 algorithm can be used for simulation purposes in order to take parameter uncertainty into account. This parameter uncertainty is caused by the finite sample length of the time series to which the model is fitted.

9.8 APPLICATIONS

9.8.1 Introduction

Three applications are presented to illustrate the advantages and usefulness of the simulation procedures presented in this chapter. The first example demonstrates that the employment of WASIM2 in simulation studies avoids bias that is due to fixed starting values. In the second application, it is shown how the model residuals from the historical data can be used in conjunction with WASIM1 for generating synthetic data. Finally, the third example demonstrates how parameter uncertainty can be incorporated into a simulation study by using WASIM3.

9.8.2 Avoidance of Bias in Simulation Studies

The *rescaled adjusted range (RAR)* and the *Hurst coefficient K* defined in [10.2.9] and [10.3.4], respectively, are two statistics that are important in problems related to the Hurst phenomenon. In Chapter 10 the controversies surrounding the *Hurst phenomenon* are presented and it is demonstrated that ARMA processes are superior to fractional Gaussian noise models for explaining the Hurst phenomenon as well as modelling annual hydrological time series. In particular, it is shown in Section 10.6 that ARMA models statistically preserve the historical RAR or equivalently K. Accordingly, ARMA models are important tools for utilization in hydrological studies.

If the underlying process is an ARMA model, it can be shown theoretically that the RAR is a function only of the sample size and the AR and MA parameters (Hipel, 1975, Appendix B). In Section 10.6, it is demonstrated how to obtain the *empirical cumulative distribution function (ECDF)* for the RAR when the generating process is a specified stochastic model. In particular, consider the ECDF for a Markov model (i.e. ARMA(1,0) process) with the AR parameter having a value of 0.7. When the WASIM2 technique is employed to generate 10,000 sequences of length 30, the value of the 0.95 quantile for the ECDF of the RAR is 12.15. The 95% confidence interval for this value is calculated to be from 12.09 to 12.19 (see Conover (1971, p. 111) for the method to calculate the confidence interval for a quantile).

If random realizations of the stochastic process are not utilized as starting values, systematic bias can be introduced into a simulation study such as the development of the ECDF for the RAR. For the Markov model with an AR parameter having a magnitude of 0.7, 10,000 sequences of length 30 were generated and for each sequence the mean value of zero was used as a starting value. In addition, exactly the same disturbances that were utilized in the simulation study using WASIM2, were employed for the biased study. The value of the 0.95 quantile for the biased ECDF of the RAR is 12.01. The 95% confidence interval for this quantile value is from 11.97 to 12.05. Notice that the confidence interval for the biased result does not intersect with the corresponding interval for the unbiased study. Consequently, fixed initial values should not be used in the development of the ECDF for a specified statistic and generating mechanism.

9.8.3 Simulation Studies Using the Historical Disturbances

When using WASIM1, it is not necessary to assume that the model residuals are NID$(0,\sigma_a^2)$. In fact, it is not required to determine any theoretical distribution for the disturbances to follow. Rather, in certain situations it may be advantageous to use the residuals from the historical data to form an empirical distribution for generating the innovations. For example, when a relatively large sample is available, it may be desirable to use the empirical distribution of the residuals for simulation studies, no matter what theoretical distribution the empirical results may most closely resemble. In other instances, it may be difficult to determine which theoretical distribution to fit to the disturbances and, consequently, it may be profitable to employ the empirical distribution of the residuals. However, it should be pointed out that when the historical disturbances are employed, it is not possible to have a generated disturbance that is more extreme than any of the calculated residuals. Nevertheless, because of the form of the difference equation for an ARMA model in [9.3.1] that is fit to correlated data, it is possible that values of the generated data may be more extreme than those in the given time series.

A riverflow time series is considered to demonstrate how the empirical distribution for the residuals can be used in practice. The average annual flows of the Gota River in Sweden are available from 1807 to 1957 in a paper by Yevjevich (1963). A model is fit to this data by following the identification, estimation, and diagnostic check stages of model construction presented in detail in Part III of the book. The identification stage reveals that it may be appropriate to estimate an AR model of order two. The parameter estimates and corresponding SE's listed in Table 9.8.1 were calculated using the unconditional sum of squares method referred to in Section 6.2.3. At the estimation stage, the white noise residuals $\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{150}\}$ are determined using the backforecasting technique of Box and Jenkins (1976, Ch. 7). Diagnostic checks performed on the residuals confirm that the modelling assumptions are satisfied. In particular, by calculating confidence limits for the residual autocorrelation function using the technique of Section 7.3.2, the residuals are shown to be white noise.

Table 9.8.1. Parameter estimates for an ARMA(2,0) model fit to the Gota River data.

Parameters	MLE's	SE's
ϕ_1	0.591	0.079
ϕ_2	-0.274	0.078

To obtain synthetic data using the Gota model, the WASIM1 method is employed and the white noise terms are chosen by selecting at random an element of the set $\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{150}\}$. After one of the historical innovations is utilized, it is put back into the set of historical disturbances. Therefore, selection is done with replacement and this method is equivalent to using the empirical distribution of the residuals for the generation of the white noise terms.

As an example of a simulation study using the Gota model, consider the development of the ECDF for the Hurst coefficient K. The historical disturbances and the WASIM1 technique are used to generate 10,000 sequences, where each sequence contains 150 values. By calculating K for each of the 10,000 traces, the ECDF for K can be obtained as shown in Table 9.8.2 for a series length which is the same as the historical time series.

The historical value of K for the Gota River is calculated to be 0.689. Notice that the observed K value does not lie in the tails of the ECDF for K in Table 9.8.2. The probability that K for the Gota model is greater than the historical K is 0.281. In Section 10.6, this procedure is applied to 23 natural time series and by invoking a particular statistical test, it is demonstrated that ARMA models do statistically preserve the Hurst coefficient K, or equivalently the RAR.

Table 9.8.2. ECDF of K for the Gota model.

Quantiles	Values of K for Empirical White Noise
0.025	0.556
0.050	0.571
0.100	0.590
0.200	0.613
0.300	0.630
0.400	0.645
0.500	0.658
0.600	0.671
0.700	0.686
0.800	0.703
0.900	0.725
0.950	0.744
0.975	0.757

9.8.4 Parameter Uncertainty in Simulation Experiments

An average annual riverflow series having a length of 96 years is modelled to show how parameter uncertainty can be brought into a practical simulation study. The yearly riverflows of the Mississippi River at St. Louis, are available from 1861 to 1957 in an article by Yevjevich (1963). By following the three stages of model development, the best process for modelling the Mississippi flows is found to be an ARMA(0,1) model. The MLE for the MA parameter θ_1 is -0.306 with a SE of 0.097.

By using WASIM1 (or equivalently WASIM2), the Mississippi model is employed to generate 10,000 series of length 96. The RAR is calculated for each of the 10,000 traces. The expected value or mean of the RAR for the 10,000 series is 13.439 with a standard deviation of 0.030.

The Mississippi model is used with WASIM3 to generate another 10,000 series of length 96. The innovations are different than those used for the simulation study with WASIM1. For each trace of length 96, the value of the MA parameter used in WASIM3 is determined by the equation

$$\theta_1 = -0.306 + 0.097\varepsilon_t \qquad [9.8.1]$$

where $t = 1,2,3,\ldots,10{,}000$; $\varepsilon_t \sim NID(0,1)$. Because the RAR is not a function of the mean level of the process or the variance of the model residuals, it is only necessary to randomly vary the MA parameter for this particular simulation study. The expected value of the RAR for the 10,000 synthetic data sets is 13.443 with a standard deviation of 0.031. A comparison of the results for the simulation experiment using a constant MA parameter with those utilizing a varying model parameter, reveals that there is no significant difference between the two expected values of the RAR. Hence, for this particular study, parameter uncertainty is not a crucial factor.

9.9 CONCLUSIONS

Improved simulation procedures are available for generating synthetic traces from ARMA and ARIMA models. Because random realizations of the underlying stochastic process are used as starting values, bias is not introduced into the simulated sequences. Furthermore, these techniques can be used in conjunction with models containing differencing operators or data that has been transformed by a Box-Cox transformation. The overall procedure for utilizing the simulation techniques are depicted in Figure 9.1.1 while detailed explanations are presented in Sections 9.2 to 9.7. Three representative applications of the simulation methods are given in Section 9.8.

If the WASIM1 method of Section 9.3 is utilized, it is not necessary that the distribution of the residuals be Gaussian. As shown by an example in Section 9.8.3, the empirical distribution of the residuals can be used for generation purposes. In addition, WASIM1 is exact for a pure MA process. On the other hand, the WASIM2 technique of Section 9.4 is an exact simulation procedure for any ARMA model. The only restriction with WASIM2 is that the residuals are NID($0,\sigma_a^2$).

When incorporating parameter uncertainty into a simulation study, the WASIM3 procedure of Section 9.7 is the proper method to implement. If it is deemed necessary to consider parameter uncertainty in reservoir design, one can employ the algorithm given in Section 9.7.3 for linking WASIM3 with the design problem. As discussed in Section 9.7.4, to circumvent difficulties with model uncertainty, it is recommended that a proper ARMA model be fit to the given data set by following three stages of model development presented in Part III.

In Section 9.5, it is explained how the simulation techniques can be used with integrated models. The simulation methods can be extended for use with the three types of seasonal models of Part VI. For example, by writing the seasonal ARIMA model in the unfactored form shown in [12.2.11], one can directly employ the simulation methods of Chapter 9.

As explained in Section 9.7.3, simulation can be used in reservoir design. Simulation can also be employed for studying the theoretical properties of a given type of stochastic model. In Section 10.6, simulation is employed to demonstrate that ARMA models preserve statistically two important historical statistics called the Hurst coefficient and the rescaled adjusted range (also see Section 9.8). This forms the basis for the explanation of what is called the Hurst phenomenon. The Hurst phenomenon and related developments in long memory modelling are presented in Chapters 10 and 11, respectively, in Part V of the book.

PROBLEMS

9.1 Explain why it is not "completely correct" to employ statistical tests for checking the statistical properties of sequences generated by a random number generator.

9.2 Briefly outline at least two statistical tests for determining whether or not a random number generator produces independent observations. List the relative merits and disadvantages of these tests.

9.3 Briefly describe at least two statistical tests for checking if a random number generator produces uniformly distributed variables. Compare the relative advantages of these tests.

9.4 In Section 9.2.3, a universal random variable generator is presented for transforming independent uniform random variables to independent random variables following any required distribution. Prove that this algorithm is correct.

9.5

(a) Provide a numerical example to demonstrate how the mixed linear congruential random number generator works.

(b) Use a numerical illustration to explain how the steps of the multiplicative linear congruential random number generator are carried out.

9.6 Describe detailed guidelines regarding the choice of the coefficients a, c and m in the linear congruential random number generator in [9.2.2].

9.7 In Section 9.2.3, references are given for the Marsaglia-Bray algorithm which can be used to generate random variables that are NID(0,1). After referring to these references, describe the main steps in this algorithm. Discuss the main advantages and drawbacks of the Marsaglia-Bray algorithm.

9.8 Three algorithms for generating NID(0,1) random variables from independent uniformly distributed random variables are the central-limit algorithm, Teichroew method and Box-Muller generator (see, for example, Knuth (1969)). Briefly describe how each algorithm works and point out any overlap in the techniques. Compare their relative advantages and disadvantages from both theoretical and computational viewpoints.

9.9 Describe an approach for generating independent gamma random variables.

9.10 Draw a flow chart to outline how one can employ simulation in reservoir design.

9.11 Suppose that one wishes to simulate 10 values from an ARMA(2,2) model using WASIM1. Write down how each of these 10 values are calculated using the WASIM1 algorithm.

9.12

(a) Suppose that one wishes to use WASIM2 to simulate 10 values from an ARMA(2,1) model. Show all the calculations for generating this data.

(b) Prove that the random starting values in part (a) are from an ARMA(2,1) process.

9.13 Refer to appropriate references for explaining how to simulate from a multivariate normal distribution. Describe in detail how this is done within the WASIM3 algorithm.

9.14 Suppose that one wishes to simulate 10 values of a series to which an ARIMA(1,1,1) model was fitted to the square roots of the observations. Write down all of the detailed calculations for producing simulated values in the untransformed domain.

9.15

(a) Using annual time series from your field of interest, carry out a simulation experiment to demonstrate that ARMA models statistically preserve the historical autocorrelation function at lag 1.

(b) Incorporate parameter uncertainty into the simulation experiment executed in part (a).

REFERENCES

DATA SET

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper no. 1, Colorado State University, Fort Collins, Colorado.

GENERATING INDEPENDENT RANDOM VARIABLES

Atkinson, A. C. and Pearce, M. C. (1976). The computer generation of beta, gamma and normal random variables. *Journal of the Royal Statistical Society, Series A*, 139(4):431-448.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29:610-611.

Dieter, V. (1972). Statistical interdependence of pseudo-random numbers generated by the linear congruential method. In Zaremba, S., editor, *Applications of Number Theory to Numerical Analysis*, pages 289-318. Academic Press, New York.

Hill, I. D. (1976). Algorithm AS100, normal-Johnson and Johnson-normal transformations. *Applied Statistics*, 25(2):190-192.

Hill, I. D., Hill, R. and Holder, R. L. (1976). Algorithm AS99, fitting Johnson curves by moments. *Applied Statistics*, 25(2):180-189.

Hull, T. and Dobell, A. (1962). Random number generators. *SIAM Review*, 4:230-254.

Janson, B. (1966). *Random Number Generators*. Victor Pettersons, Bokindustri, Akiebolag, Stockholm.

Johnson, N. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36:149-176.

Knuth, D. E. (1969). *The Art of Computer Programming*. Addison-Wesley, Reading, Massachusetts.

Lehmer, D. H. (1949). Mathematical methods in large scale computing units. In *Proceedings of the Symposium on Large Scale Digital Calculating Machinery*, pages 141-146. Harvard University Press.

Marsaglia, G. (1972). The structure of linear congruential sequences. In Zaremba, S., editor, *Applications of Number Theory to Numerical Analysis*, pages 249-286. Academic Press, New York.

Marsaglia, G. and Bray, T. A. (1964). A convenient method for generating normal variables. *SIAM Review*, 6:260-264.

Marsaglia, G., MacLaren, G. M. M. and Bray, T. A. (1964). A fast procedure for generating normal variables. *Communications of the Association of Computing Machinery*, 7:4-10.

von Neumann, J. (1951). Various techniques used in connection with random digits. *NBS Applied Mathematics Series*, (12):36-38.

Weckmann, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, 31:188-190.

Yakowitz, S. J. (1977). *Computational Probability and Simulation*. Addison-Wesley, Reading, Massachusetts.

HYDROLOGICAL SIMULATION RESEARCH

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised edition.

Delleur, J. W., Tao, P. C., and Kavvas, M. L. (1976). An evaluation of the practicality and complexity of some rainfall and runoff time series models. *Water Resources Research*, 12(5):953-970.

Hipel, K. W. (1975). Contemporary Box-Jenkins Modelling in Water Resources. PhD thesis, University of Waterloo, Waterloo, Ontario.

Hipel, K. W. and McLeod, A. I. (1978). Preservation of the rescaled adjusted range, 2, simulation studies using Box-Jenkins models. *Water Resources Research*, 14(3):509-516.

Kottegoda, N. T. (1987). Fitting Johnson S_B curve by the method of maximum likelihood to annual maximum daily rainfalls. *Water Resources Research,* 23(4):728-732.

McLeod, A. I. and Hipel, K. W. (1978). Simulation procedures for Box-Jenkins models. *Water Resources Research,* 14(5):969-975.

Moran, P. A. P. (1959). *On the Theory of Storage.* Methuen, London.

Sangal, B. P. and Biswas, A. K. (1970). The three-parameter lognormal distribution and its application in hydrology. *Water Resources Research,* 6(2):505-515.

Stedinger, J. R. (1980). Fitting log normal distributions to hydrologic data. *Water Resources Research,* 16:481-490.

Vogel, R. M. and Stedinger, J. R. (1988). The value of stochastic streamflow models in overyear reservoir design applications. *Water Resources Research,* 24(9):1483-1490.

INITIAL VALUES

Brown, T. J. and Hardin, C. (1973). A note on Kendall's autoregressive series. *Journal of Applied Probability, 10:475-478.*

Copas, J. B. (1966). Monte Carlo results for estimation of a stable Markov time series. *Journal of the Royal Statistical Society, Series A,* 129:110-116.

STATISTICS AND NUMERICAL ANALYSIS

Conover, W. J. (1971). *Practical Nonparametric Statistics.* John Wiley, New York.

Healy, M. J. R. (1968). Algorithm AS6, triangular decomposition of a symmetric matrix. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 17:195-197.

Hornbeck, R. W. (1975). *Numerical Methods.* Quantum Publishers, New York.

Ralston, A. (1965). *A First Course in Numerical Analysis.* McGraw-Hill, New York.

PART V

LONG MEMORY MODELLING

The **Hurst phenomenon** created one of the most interesting, controversial and long-lasting scientific debates ever to arise in the field of hydrology. The genesis of the Hurst phenomenon took place over forty years ago in Egypt. Just after World War II, a British scientist by the name of **Harold Edwin Hurst** became deeply involved in studying how the Nile River could be optimally controlled and utilized for the benefit of both Egypt and Sudan. As Director-General of the Physical Department in the Ministry of Public Works in Cairo, Egypt, Hurst was particularly interested in the **long-term storage** requirements of the Nile River. In addition to annual riverflow series, Hurst analyzed a wide variety of other yearly geophysical time series in order to examine the statistical properties of some specific statistics that are closely related to long term storage. These statistical studies led Hurst to develop an empirical law upon which the definition of the Hurst phenomenon is based.

The fact that the Hurst phenomenon arose from scientific work carried out in Egypt provided the controversy with an aura of mystery and intrigue. Was the Hurst phenomenon more difficult to solve than the riddle of the Sphinx? Indeed, a range of explanations has been put forward for solving the Hurst phenomenon. Furthermore, in the process of studying the Hurst phenomenon, many original contributions have been made to the fields of hydrology and statistics.

In Chapter 10, the **Hurst phenomenon** is defined and both theoretical and empirical work related to this phenomenon are described. One spinoff from research connected to Hurst's work is the development of a stochastic model called **fractional Gaussian noise (FGN)**. This model possesses **long memory** (see Section 2.5.3) and was designed for furnishing an explanation to the Hurst phenomenon. As demonstrated in Chapter 10, this long memory model fails to solve the Hurst riddle. Nevertheless, the introduction of FGN into the field of hydrology initiated major theoretical and practical developments in long memory modelling by not only hydrologists but also by statisticians and economists. Probably the most flexible and comprehensive type of long memory model is the **fractional autoregressive-moving average or FARMA model** presented in Chapter 11. In fact, the FARMA family of models is a direct extension of the ARMA class of models defined in Chapter 3.

If FGN modelling cannot provide a reasonable solution to the Hurst phenomenon, then wherein lies the answer? The solution to Hurst's riddle is put forward in Section 10.6 of the next chapter. Simulation experiments demonstrate that when the most appropriate ARMA models are fitted to a wide variety of annual natural time series, a statistic called **the Hurst coefficient is "statistically preserved" by the calibrated ARMA models.** Therefore, although the Hurst coefficient and other related statistics are not directly incorporated as model parameters in the design of an ARMA model, these statistics can still be indirectly accounted for or modelled by ARMA models.

CHAPTER 10

THE HURST PHENOMENON

AND

FRACTIONAL GAUSSIAN NOISE

10.1 INTRODUCTION

Since the original empirical studies of Hurst (1951), the *Hurst phenomenon* has caused extensive research with accompanying academic controversies right up to the present time. The objectives of this chapter are to review and appraise research related to Hurst's work and demonstrate how the Hurst phenomenon can be explained. The views presented in this chapter, as well as research by Hipel (1975), McLeod and Hipel (1978a) and Hipel and McLeod (1978a, 1978b), constitute a fresh approach to the study of the Hurst phenomenon and the related problem of the *preservation of historical statistics* by stochastic models.

In Section 10.2, some statistics related to long term storage requirements of a reservoir are defined using the idea of a *cumulated range.* Subsequently, the various types of Hurst coefficients that have been developed for use in formulae involving the *rescaled adjusted range (RAR)* are given and compared in Section 10.3.1. Because of the flexible statistical properties of the RAR, it is suggested that this is the Hurst statistic of primary concern in water resource applications related to storage.

The roles of both identically independently distributed (IID) variables and correlated random variables for explaining problems related to the Hurst phenomenon are thoroughly investigated in Sections 10.3.2 and 10.3.3, respectively. Simulation studies are used to demonstrate that the RAR is nearly independent of the type of underlying distribution of the random variables and is also a function of the sample size. Of particular importance for correlated processes are stochastic models that can be easily fitted to natural time series and at the same time retain relevant historical statistical characteristics of the data such as the RAR and other related statistics. The ARMA models of Chapter 3 constitute one family of stochastic or time series models which possess the potential for continued extensive utilization in hydrology. The *fractional Gaussian noise (FGN) model* of Section 10.4 is a process that was developed mainly within the hydrological literature (Mandelbrot and Wallis, 1968, 1969a to e) as a means for possibly accounting for the Hurst phenomenon. Although some of the inherent drawbacks of this model are discussed, significant contributions are formulated toward the further statistical maturity of the FGN model in Section 10.4.

As explained in Part III, when any type of time series model is being fitted to a given time series, it is recommended to follow the identification, estimation and diagnostic check stages of model construction. Within Section 10.4, useful model building techniques are presented for allowing FGN models to be applied properly to data sets. More specifically, in Section 10.4.3 an efficient maximum likelihood estimation (MLE) procedure is derived for use at the estimation stage. Simulation studies reveal that the MLE approach is superior to a previous estimation method. A technique for calculating the model residuals is given in Section 10.4.4, so that the statistical properties of the residuals can be tested by specified diagnostic checks. If, for example, the residuals fail to pass the whiteness criterion, another type of model should be chosen in order to satisfy this important modelling assumption. Next, a procedure is presented in Section 10.4.5 for calculating minimum mean square error (MMSE) forecasts for a FGN model. Following this, an exact simulation procedure is given in Section 10.4.6 for simulating FGN. This new simulation method eliminates the need for approximating FGN by other types of stochastic processes.

The FGN model is an example of what is called a long memory process defined in Section 2.5.3. On the other hand, the ARMA models of Chapter 3 possess short memory. For discriminating between the long memory FGN models and the short memory ARMA models, the Akaike information criterion (AIC) defined in Section 6.3.2 can be employed. For the six annual riverflow time series considered in Section 10.4.7, the AIC selects the ARMA model in preference to the FGN model in each case.

To investigate statistical properties of the RAR and the *Hurst coefficient K*, simulation experiments are carried out in Section 10.5. Within Section 10.5.2, simulation studies are executed using white noise while in Section 10.5.3 the simulation studies involve synthetic data generated from both long and short memory models.

A major challenge in stochastic hydrology is to determine time series models that preserve important historical statistics such as the RAR, or equivalently, the Hurst coefficient K. By following the identification, estimation, and diagnostic check stages of model development, ARMA models are determined for 23 geophysical time series in Section 10.6. Simulation studies are then performed to determine the small sample *empirical cumulative distribution function (ECDF)* of the RAR or K for various ARMA models. The ECDF for these statistics is shown to be a function of the time series length N and the parameter values of the specific ARMA model being considered. Furthermore, it is possible to determine as accurately as desired the distribution of the RAR or K. A theorem is given to obtain confidence intervals for the ECDF in order to guarantee a prescribed precision. Then it is shown by utilizing simulation results and a given statistical test that ARMA models do preserve the observed RAR or K of the 23 geophysical time series. Consequently, *ARMA models provide an explanation for the Hurst phenomenon*. Finally, various estimates for the Hurst coefficient are estimated and compared in Section 10.7 for the 23 given time series.

The FGN model defined in this chapter is one example of a long memory model. Another example is the mixed Gamma ARMA(1,1) model proposed by Sim (1987). A flexible class of long memory models based on sound theoretical foundations is the *fractional autoregressive-moving average or FARMA family of models*. In reality, FARMA models constitute direct extensions of ARMA and ARIMA models. Within Chapter 11, FARMA models are defined and model construction techniques are presented.

10.2 DEFINITIONS

The definitions presented in this section reflect long term storage requirements of reservoirs and are needed for explaining the Hurst phenomenon in Section 10.3.1.

Consider a time series $z_1, z_2, \ldots, z_N$. Define the *kth general partial sum* as

$$S'_k = S'_{k-1} + (z_k - \alpha \bar{z}_N) = \sum_{i=1}^{k} z_i - \alpha k \bar{z}_N \,, \quad k = 1,2,\ldots,N \qquad [10.2.1]$$

where S'_0 equals 0, $\bar{z}_N = \frac{1}{N}\sum_{i=1}^{N} z_i$ is the mean of the first N terms of the series, and α is a constant satisfying $0 \leq \alpha \leq 1$. The *general (cumulative) range* R'_N is defined as

$$R'_N = M'_N - m'_N \qquad [10.2.2]$$

where $M'_N = \max(0, S'_1, S'_2, \ldots, S'_N)$ is the *general surplus*, and $m'_N = \min(0, S'_1, S'_2, \ldots, S'_N)$ is the *general deficit*. Thus, R'_N is the range of cumulative departures of the random variables $z_1, z_2, \ldots, z_N$, from $\alpha \bar{z}_N$. When random variables such as $z_1, z_2, \ldots, z_N$, are employed in summation operations, they are often referred to as *summands*. The *rescaled general range* R'_N is given as

$$\bar{R}'_N = R'_N / D'_N \qquad [10.2.3]$$

where $D'_N = N^{-1/2}\left[\sum_{i=1}^{N}(z_i - \alpha \bar{z}_N)^2\right]^{1/2}$ is the *general standard deviation*.

The constant α can be thought of as an *adjustment factor*, or in storage theory, it can be interpreted as the degree of development of reservoir design. Two special cases for α are of particular importance in water resources. For $\alpha = 0$ (no adjustment) the kth general partial sum S'_k is replaced by the *crude partial sum* S_k, which is defined by

$$S_k = S_{k-1} + z_k = \sum_{i=1}^{k} z_i \,, \quad k = 1,2,\ldots,N \qquad [10.2.4]$$

where $S_0 = 0$. The *crude range* R_N is defined analogous to R'_N as

$$R_N = M_N - m_N \qquad [10.2.5]$$

where $M_N = \max(0, S_1, S_2, \ldots, S_N)$ is the *crude surplus*, and $m_N = \min(0, S_1, S_2, \ldots, S_N)$ is the *crude deficit*. Similarly, the *rescaled crude range* is

$$\bar{R}_N = R_N / D_n \qquad [10.2.6]$$

where $D_N = N^{-1/2}\left[\sum_{i=1}^{N} z_i^2\right]^{1/2}$ is the *crude deviation*.

When $\alpha = 1$ (maximum adjustment or development), the *kth adjusted partial sum* S^*_k is given by

$$S^*_k = S^*_{k-1} + (z_k - \bar{z}_N) = \sum_{i=1}^{k} z_i - k\bar{z}_N \,, \quad k = 1,2,\ldots,N \qquad [10.2.7]$$

where $S^*_0 = 0$ and $S^*_N = 0$. The *adjusted range* R^*_N is defined as

$$R^*_N = M^*_N - m^*_N \qquad [10.2.8]$$

where $M^*_N = \max(0, S^*_1, S^*_2, \ldots, S^*_N)$ is the *adjusted surplus*, and $m^*_N = \min(0, S^*_1, S^*_2, \ldots, S^*_N)$ is the *adjusted deficit*. Finally, the *rescaled adjusted range* is

$$\bar{R}^*_N = R^*_N / D^*_N \qquad [10.2.9]$$

where $D^*_N = N^{-1/2}\left[\sum_{i=1}^{N}(z_i - \bar{z}_N)^2\right]^{1/2}$ is the *sample standard deviation*. Figure 10.2.1 graphically illustrates the concepts of S^*_k, M^*_N, m^*_N, and R^*_N.

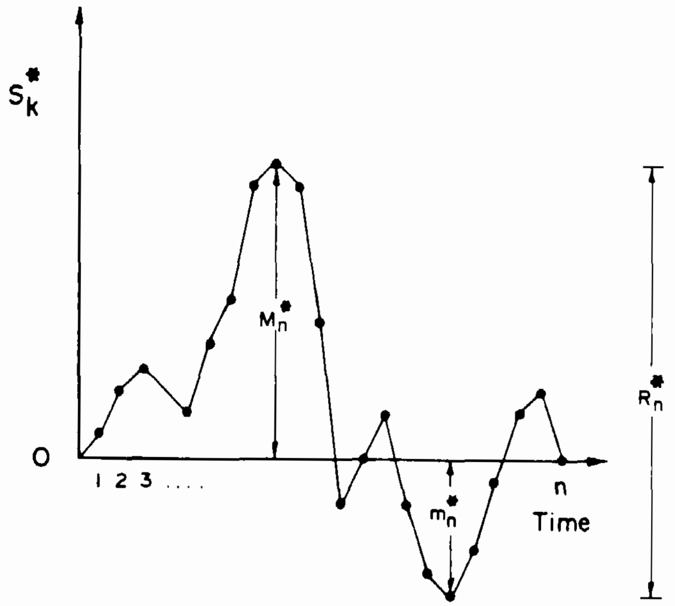


Figure 10.2.1. Adjusted range.

The statistics described in this section are extremely useful in *reservoir design*. If the z_t are average annual volumes of riverflow, then $\sum_{i=1}^{k} z_i$ is the inflow into a reservoir in k years, and $\alpha k \bar{z}_N$ is the outflow at a level of development α. The S'_k in [10.2.1] represents the *storage* after k years. Also, R'_N is the minimum reservoir capacity required to satisfy a constant draft of $\alpha \bar{z}_N$ without experiencing shortages or spills over the period spanned by the inflow sequence $z_1, z_2, \ldots, z_N$. When $\alpha = 1$, the water in the river would be used to its full potential.

The time series $z_1, z_2, \ldots, z_N$ is said to be *covariance stationary* (see Section 2.4.2) if the mean

$$\mu = E[z_t] \qquad [10.2.10]$$

and the theoretical *autocovariance function (ACF)*

$$\gamma_k = E[(z_t - \mu)(z_{t-k} - \mu)] \qquad [10.2.11]$$

both exist and do not depend on t. The statistical properties of any covariance stationary Gaussian time series are completely determined by its mean μ, variance γ_0, and theoretical *autocorrelation function (ACF)*,

$$\rho = \gamma_k / \gamma_0 \qquad [10.2.12]$$

The physical interpretations of the stationary assumptions are discussed in Section 2.4 and also by Klemes (1974).

Often it seems reasonable to assume that recent values of a time series contain more information about the present and future than those in the remote past. Accordingly, it is assumed that the theoretical ACVF is *summable* as defined by (Brillinger, 1975)

$$M = \sum_{k=-\infty}^{\infty} |\gamma_k| < \infty \qquad [10.2.13]$$

As is also mentioned in Section 2.5.3, a covariance stationary time series model is said to have a *short or a long memory* according to whether the theoretical ACVF (or equivalently the theoretical ACF) is summable. Thus, the FGN model has a long memory (for the model parameter H in the range $0.5 < H < 1$), whereas the ARMA models have a short memory. For a specified range of a model parameter d, the FARMA models of Chapter 11 also possess long memory.

10.3 HISTORICAL RESEARCH

10.3.1 The Hurst Phenomenon and Hurst Coefficients

Hurst (1951, 1956) stimulated interest in the RAR statistic by his studies of long-term storage requirements for the Nile River. On the basis of a study of 690 annual time series comprising streamflow, river and lake levels, precipitation, temperature, pressure, tree ring, mud varve, sunspot and wheat price records, Hurst implied that $\bar{R}^*_N$ varies with N as

$$\bar{R}^*_N \propto N^h \qquad [10.3.1]$$

where h is the *generalized Hurst coefficient.* The above equation can be written in the general form

$$\bar{R}^*_N = aN^h \qquad [10.3.2]$$

where a is a coefficient that is not a function of N. It should be noted that Hurst did not explicitly state the generalized Hurst law of [10.3.2] in his research papers. However, by choosing the coefficient a to have a value of $(1/2)^h$, Hurst in effect estimated h by the *Hurst coefficient K* in the empirical equation

$$\bar{R}^*_N = (N/2)^K \qquad [10.3.3]$$

By taking logarithms of [10.3.3], an explicit relationship for K is then

$$K = \frac{\log\bar{R}^*_N}{\log N - \log 2} = \frac{\log R^*_N - \log D^*_N}{\log N - \log 2} \qquad [10.3.4]$$

Employing series that varied in length from 30 to 2000 years, Hurst found K to range from 0.46 to 0.96 with a mean of 0.73 and a standard deviation of 0.09.

Assuming a normally independently distributed (NID) process, Hurst (1951) utilized some coin-tossing experiments to develop the theoretical asymptotic relationship for the expected value of the adjusted range as

$$E[R^*_N] = (\pi N\gamma_0/2)^{1/2}$$

or

$$E[R^*_N]/(\gamma_0)^{1/2} = 1.2533N^{1/2} \qquad [10.3.5]$$

Using the theory of Brownian motion, Feller (1951) rigorously established the above asymptotic formula for any sequence of IID random variables possessing finite variance. It follows from a standard convergence theorem in probability theory (Rao, 1973, p. 122) that for large N,

$$E(\bar{R}^*_N) = 1.2533N^{1/2} \qquad [10.3.6]$$

Even though Hurst studied the RAR for small N and not for the adjusted range, the form of [10.3.5] prompted him to use K in [10.3.4] as an estimate of h and also to assume K to be constant over time. However, for 690 geophysical time series, Hurst found K to have an average of 0.73, while the asymptotic, or limiting, value of K given by [10.3.6] is 0.5. This discrepancy is referred to as the *Hurst phenomenon*. The search for a reasonable explanation of the Hurst phenomenon and the need for methods whereby the statistics related to Hurst's work can be incorporated into mathematical models have intrigued researchers for decades.

In addition to K, other estimates of the generalized Hurst coefficient h in [10.3.2] have been formulated. Based upon the structure of [10.3.6], Gomide (1975, 1978) suggested estimating h by the YH that is given in the following equation:

$$\bar{R}^*_N = 1.2533N^{YH} \qquad [10.3.7]$$

The average value of YH for the 690 series considered by Hurst is 0.57 rather than 0.73.

Siddiqui (1976) proposed a method of evaluating h if the underlying process is assumed to be an ARMA process. The estimate of Siddiqui is the result of a comparison between an asymptotic result for calculating $E(\bar{R}^*_N)$ for ARMA processes and the form of [10.3.2]. Siddiqui's estimate of h and the statistic YH of Gomide (1975, 1978) are calculated in Section 10.7 for the 23 geophysical time series considered in Section 10.6. Appropriate conclusions are drawn regarding the behaviour of these statistics in relationship to K and whether they exhibit the Hurst phenomenon. For the case of a white noise process, Siddiqui's estimate of h is identical with Gomide's statistic YH in [10.3.7].

For NID random variables, Anis and Lloyd (1976) have suggested a specific estimate of h that is a function of the sample size. By taking logarithms of [10.3.2] for the expected value of the RAR, the following equation is obtained.

$$\log E[\bar{R}^*_N] = \log a + h\log N \qquad [10.3.8]$$

Anis and Lloyd (1976) defined the local Hurst exponent $h(N)$ as the derivative

$$h(N) = \partial(\log E[\bar{R}^*_N])/\partial(\log N) \qquad [10.3.9]$$

The exponent $h(N)$ can be tabulated approximately from the equation

$$h(N) = \frac{\log E[\bar{R}^*_{N+1}] - \log E[\bar{R}^*_{N-1}]}{\log(N+1) - \log(N-1)} \qquad [10.3.10]$$

where $E(\bar{R}^*_N)$ is calculated exactly by using the formula of Anis and Lloyd (1976) that is also given in [10.3.15]. It should be noted that previously Salas-La Cruz and Boes (1974) had defined an exponent similar to $h(N)$ for the general range where $0 \le \alpha \le 1$.

Because the entries for the expected value of the RAR on the right-hand side of [10.3.10] are calculated directly from a theoretical formula, $h(N)$ is not a function of the data and is, therefore, not a statistic. Nevertheless, it would perhaps be possible to fit some type of stochastic model to a given time series and then to derive the RAR terms in [10.3.10] by using simulation. Most likely, this type of procedure may not be a worthwhile venture, and hence $h(N)$ probably will have limited use in practical hydrological problems.

Anis and Lloyd (1976, p. 115, Table 1) list values of $h(N)$ for N ranging from 5 to 10^6. Although the magnitude of $h(N)$ asymptotically approaches 0.5 for increasing N, at lower values of N, the $h(N)$ is significantly larger than 0.5. For instance, when N possesses values of 5, 40, 100, 200, and 500, then $h(N)$ has magnitudes of 0.6762, 0.5672, 0.5429, 0.5315, and 0.5202, respectively.

In the development of an estimate for the parameter H in FGN models, Mandelbrot and Wallis (1969d) assumed a form of the Hurst law that is identical with [10.3.2]. For a given time series $z_1, z_2, \ldots, z_N$, let $\bar{R}^*_{r'}(t,r)$ denote the RAR of the subseries $z_t, z_{t+1}, \ldots, z_r$, and let $r' = r - t + 1$. When examining scatter plots (or "pox diagrams") of $\log\bar{R}^*_{r'}(t,r)$ versus $\log r'$ for a number of selected values of t and r, Mandelbrot and Wallis (1969d) were using for each subseries a Hurst law given by

$$\bar{R}^*_{r'}(t,r) = a(r')^h \qquad [10.3.11]$$

Wallis and Matalas (1970) have suggested the G Hurst estimator for estimating the parameter H in FGN models and also h in [10.3.11]. This procedure estimates h by calculating the slope of the regression of the averaged values of $\log\bar{R}^*_{r'}(t,r)$ on $\log r'$ for specified values of t and r.

When Hurst originally formulated [10.3.3] there is no doubt that he was attempting to derive an empirical law that would be valid for a wide range of geophysical phenomena. In particular, an equation such as [10.3.3] would be extremely useful for reservoir design if the phenomenon being modelled were average annual riverflows. However, the distribution of K plus the other types of Hurst exponents summarized in this section are a function of the sample size N. For example, the empirical cumulative distribution functions of K for various values of

N for certain types of ARMA processes are given in Section 10.5.3. In addition, as shown in Section 10.6, when K is estimated for 23 given geophysical time series, K seldom has exactly the same value for any given pair of data sets. Because of the aforementioned facts, the empirical law of Hurst in [10.3.3] loses much of its simplicity and also its potential for being a universal law. This inherent lack of universality of Hurst's law may be due to the fact that the general form of [10.3.3] resembles the asymptotic formula given in [10.3.6], whereas in practice it is necessary to deal with small and moderate sample sizes.

Because the RAR possesses many attractive statistical features, Hurst perhaps should have concentrated his efforts on studying the properties of $\bar{R}^*_N$ rather than those of K. The RAR statistic is independent of the magnitude of the mean level and standard deviation of a time series. If the data are modelled by an ARMA process, $\bar{R}^*_N$ is only a function of the sample size N and the autoregressive (AR) and moving average (MA) parameters and is independent of the variance of the innovations (Hipel, 1975, Appendix B). From [10.3.4] it can be seen that K is simply a transformation of $\bar{R}^*_N$ and, therefore, also possesses the aforementioned properties of the RAR. Nevertheless, the formulation of K in [10.3.4] as a function of $\bar{R}^*_N$ only introduces an unnecessary transformation and does not give K any additional advantageous statistical properties that are not already possessed by the RAR. It is therefore recommended that future research should concentrate on the RAR rather than on the various types of Hurst exponents discussed in this section.

Since the concept of the Hurst coefficient is so entrenched in the literature, it is widely quoted in the remainder of this chapter. However, the reader should be aware that the statistic of primary concern is the RAR. Even the use of the G Hurst statistic (Wallis and Matalas, 1970), which was primarily developed as an estimate for the parameter H in FGN models, is questionable. It is demonstrated later in this chapter that a MLE of H is a more efficient procedure to employ.

10.3.2 The Hurst Phenomenon and Independent Summands

Besides the results of Feller (1951), Hurst's work influenced other researchers to develop theoretical derivations for statistics related to the cumulative range. Because of the mathematical complexity in deriving theoretical formulae for the moments of statistics connected with the range, a large portion of the research was devoted to the special case of independent summands. Anis and Lloyd (1953) developed a formula for the expected value of the crude range for standard NID variates. Anis (1955) derived the variance of M_N and subsequently a method for obtaining all the moments of M_N (Anis, 1956).

Solari and Anis (1957) determined the mean and variance of the adjusted range for a finite number of NID summands. Feller (1951) had noted that the sampling properties of the adjusted range were superior to those of the crude range. The results of Solari and Anis (1957) for the variance of M^*_N substantiated the conclusion of Feller when this variance was compared to that of M_N (Anis, 1956).

Moran (1964) initiated a new line of development when he observed that the expected value of cumulative ranges could easily be derived from a combinatorial result known as Spitzer's lemma. He showed that for moderate N, distributions with very large second moments about the mean could cause the $E(M_N)$ to increase more quickly than $N^{1/2}$. This, in turn, implied that the crude range would do likewise.

For independently, stably distributed summands with the characteristic exponent ν, Boes and Salas-La Cruz (1973) showed that asymptotically

$$E(\bar{R}^*_N) \propto N^{1/\nu} \qquad [10.3.12]$$

where $1 < \nu \leq 2$. The general stable distribution with characteristic exponent ν is defined for $1 < \nu \leq 2$ in terms of its characteristic function

$$\omega(u) = E(e^{iuz_t}) \qquad [10.3.13]$$

by

$$\log\omega(u) = i\mu u + \sigma^2 |u|^\nu \left\{ 1 + i\beta(u/|u|)\tan[(\pi/2)\nu] \right\} \qquad [10.3.14]$$

where $i = (-1)^{1/2}$, μ is the location parameter for the random variable z_t; σ is the scale parameter for the random variable z_t; and β is the measure of skewness. For $\beta = 0$ and $\nu = 2$, the normal distribution is obtained. Stable distributions with characteristic exponent $1 < \nu < 2$ generate more extreme observations than the normal distribution. Granger and Orr (1972) have suggested that economic time series are best modelled by a stable distribution with characteristic exponent $1.5 < \nu < 2$. From [10.3.12] it could be suggested that a stable distribution with $\nu = 1.37$ (approximately) for geophysical time series could explain Hurst's findings. However, because for the case of stable distributions with $1 < \nu < 2$ the sample variance is not a consistent estimator of the scale parameter σ, it does not follow that [10.3.12] will hold for the RAR. In fact, simulation experiments reported later in this chapter show that the expected value of the RAR for independently stably distributed summands with characteristic exponent $\nu = 1.3$ very nearly equals the expected value of the RAR for NID summands.

All of the aforesaid research was influenced by the original work of Hurst. However, mathematicians have for a long time been investigating the crude range of independent summands independently of Hurst's empirical research. Anis and Lloyd (1976) give a brief survey of mathematical studies of the crude range. Further references can also be found in a paper by Berman (1964).

Unfortunately, none of the foregoing theoretical investigations discussed in this section have dealt with the RAR. However, for a NID process, Anis and Lloyd (1976) have successfully proved the following exact equation to be the expected value of the RAR:

$$E(\bar{R}^*_N) = \frac{\Gamma[\frac{1}{2}(N-1)]}{(\pi)^{1/2}\Gamma(\frac{1}{2}N)} \sum_{r=1}^{N-1} \frac{N-r}{r} \qquad [10.3.15]$$

10.3.3 The Hurst Phenomenon and Correlated Summands

Introduction

When Hurst (1951) theoretically derived [10.3.5] for the adjusted range, he assumed normality of the process, he developed that equation as an asymptotic relationship relationship, and he assumed independence of the time series. As was pointed out by Wallis and Matalas (1970), these three facts respectively caused the following three possible explanations of the Hurst phenomenon: (1) nonnormality of the probability distribution underlying the time series, (2) transience (i.e., N is not large enough for the Hurst coefficient to attain its limiting value of 0.5), and (3) autocorrelation due to nonindependence.

For independent summands, nonnormality of the underlying process has largely been discounted as a possible explanation of the Hurst phenomenon. If a very large sample is being considered, the asymptotic expression in [10.3.6] has been shown to be valid for IID random variables. For samples of small and moderate lengths, simulation studies later in this chapter (see Table 10.5.1) reveal that the RAR is very nearly independent of the distribution of the random variables. Because the Hurst coefficient K is definitely a function of N for independent summands (see, for example, Table 10.5.2 for the NID case), then transience constitutes a plausible explanation to Hurst's dilemma (also see Salas et al. (1979)).

For the autocorrelated case, Wallis and O'Connell (1973) correctly concluded that transience is obviously connected with the autocorrelation structure of the generating process, and, therefore, these two effects must be considered simultaneously when attempting to account for the Hurst phenomenon. As is illustrated by simulation studies in Sections 10.5 and 10.6 for ARMA models, both transience and autocorrelation form an explanation of the Hurst phenomenon. In this section, the roles of both short memory and long memory processes for explaining and modelling the Hurst phenomenon are examined.

Hurst (1951) actually conjectured that K had a value of 0.73 and not 0.5 because of persistence. This is the tendency for high values to be followed by high values and low values by low values which are referred to by Mandelbrot and Wallis (1968) as the Noah and Joseph effects, respectively. Persistence is caused by the dependence of naturally occurring time series as exhibited in their serial correlation structure. For reservoir design this means that for a given value of N the size of a reservoir that releases the mean flow each year would need to be larger than the capacity corresponding to an uncorrelated series of inflows.

Short Memory Models

Barnard (1956) and Moran (1959) observed that for the standard short memory time series models the following asymptotic formula is valid:

$$E(\bar{R}^*_N) = aN^{1/2} \qquad [10.3.16]$$

where a is a coefficient that does not depend on N. Mandelbrot and Van Ness (1968) proved that for large N, [10.3.16] holds for any short memory time series model. Siddiqui (1976) demonstrated that for any model with a summable theoretical ACVF,

$$a = \left(\frac{\pi}{2\gamma_0} \sum_{i=-\infty}^{\infty} \gamma_i \right)^{1/2} \qquad [10.3.17]$$

It has been argued by some authors that because short memory models, such as the ARMA processes, imply a limiting value of K equal to 0.5 and since the observed K in annual geophysical time series is about 0.7, short memory models are not appropriate models for synthetic streamflow generation. It should therefore be emphasized that asymptotic results are only relevant in that they provide an approximation to the exact results for the true (finite) series length.

Anis and Lloyd (1976) showed that [10.3.15] also holds exactly for symmetrically correlated normal summands. But such a time series has a long memory, since its theoretical ACVF is not summable. Because [10.3.15] is also valid for short memory NID random variables, this fact provides a counterexample to the claim of some researchers that long memory models are a necessary explanation of the Hurst phenomenon. Conversely, Klemes (1974) has shown that zero memory nonstationary models could produce the Hurst phenomenon. By simulation experiments with white noise, he varied the mean level in different manners and showed how K increased in value due to this type of nonstationarity. Klemes also demonstrated by simulation that random walks with one absorbing barrier, which often arise in natural storage systems, could cause the RAR to have certain properties related to the Hurst phenomenon.

Hurst (1957) was the first scientist to suggest that a nonstationary model in which the mean of the series was subject to random changes could account for higher values of the Hurst coefficient K and hence the Hurst phenomenon. Similar models have been studied by Klemes (1974) and Potter (1976). As generalizations of the models of Hurst (1957), Klemes (1974) and Potter (1976), the *shifting level processes* were developed by Boes and Salas (1978). Further research in shifting level processes is provided by Salas and Boes (1980), Ballerini and Boe (1985), and Smith (1988). The basic idea underlying a shifting level process is that the level of the process randomly shifts to different levels which last for random time periods as the process evolves over time.

In a four page commentary, D'Astous et al. (1979) demonstrated that the annual precipitation data employed by Potter (1976) may not justify the concept of a shifting level time series. Using simulation and the segmentation scheme suggested by Potter (1976) for isolating shifting levels, they showed that an ARMA(1,1) process can mimic this type of changing level. If the mean level of a time series changes due to known natural or human intervention, then the intervention model of Chapter 19 can be used to model the data.

Matalas and Huzzen (1967) performed statistical experiments to determine whether K is preserved by Markov models. For values of the lag 1 autocorrelation coefficient ρ_1 ranging from 0 to 0.9, they calculated the $E(K)$ based upon 10^4 simulations for particular values of N and ρ_1. For values of N and ρ_1, compatible with what occurs in annual riverflows if those flows are assumed Markov, they found K to have an average of about 0.7. Because a mean of approximately 0.7 for K occurs in natural time series, they implied that perhaps the small sample properties of K are preserved by a Markov model. Nevertheless, a later simulation study of Wallis and Matalas (1970) suggested that the observed sample lag 1 autocorrelations for flows in the Potomac River basin were too low for a first order AR process adequately to preserve the Hurst K. However, a Markov model may not necessarily be the best short memory model to fit to a given time series. Rather, it is recommended to select the proper ARMA model by adhering to the identification, estimation, and diagnostic check stages of model construction, as explained in Part III of this book. In some cases, the appropriate model may indeed be a Markov model. In Section 10.6, it is demonstrated that, for 23 geophysical time series ranging in length from $N = 96$ to $N = 1164$, properly fitted ARMA models do adequately preserve K.

Several other authors have also suggested that short memory models may preserve K. Gomide (1975) has completed further simulation studies of the RAR for Markov models. O'Connell (1974a,b) advocated employing an ARMA(1,1) model to approximate the long memory FGN model and thereby perhaps to preserve K. To accomplish this, the AR parameter must have a value close to unity, so that the ACF of the process will attenuate slowly and hence approximate the theoretical ACF of the FGN process. In practice, this approach may not be viable. The proper ARMA model that is fit to the data may not be ARMA(1,1), and even if it is ARMA(1,1), an efficient MLE of the parameters may not produce an estimate of the AR parameter that is close to 1. This parameter estimation problem is acknowledged by O'Connell (1976). In addition, it is no longer necessary to approximate FGN by a short memory model such as an ARMA(1,1) model because as is shown in Section 10.4.6 it is now possible to simulate FGN exactly.

Long Memory Models

A long memory model known as FGN was introduced into the hydrological literature by Mandelbrot and Wallis (1968, 1969a to e) to explain the Hurst phenomenon. In Section 10.4, the FGN model is defined and new developments in FGN modelling are presented. Other research on stochastic processes related to FGN is given by authors such as Taqqu (1979) and Cox (1984). In Chapter 11, the theory and practice of the long memory FARMA class of models is presented.

10.4 FRACTIONAL GAUSSIAN NOISE

10.4.1 Introduction

The connection between FGN and Hurst's law is the parameter H in FGN that is often estimated using the Hurst coefficient K in [10.3.4]. The FGN model was first proposed by Mandelbrot (1965), and a mathematical derivation was given by Mandelbrot and Van Ness (1968) and Mandelbrot and Wallis (1969c). The literature concerning the FGN model has been summarized by authors such as Wallis and O'Connell (1973), O'Connell (1974b, Ch. 2), Hipel (1975), Lawrance and Kottegoda (1977) and McLeod and Hipel (1978a). Consequently, only the main historical points of practical interest are discussed in Section 10.4.2. Following a brief historical description and definition of FGN in the next section, new advancements are presented. These include efficient parameter estimation, model diagnostic checking, forecasting, and exact simulation. In an application section, FGN models are compared to ARMA models when both types of models are fitted to six average annual riverflow series.

10.4.2 Definition of FGN

In the development of FGN processes, Mandelbrot (1965) considered a continuous time process $B_H(t)$ that satisfies the self-similarity property such that for all τ and $\varepsilon > 0$, $B_H(t + \tau) - B_H(t)$ has exactly the same distribution as $[B_H(t + \tau\varepsilon) - B_H(t)]/\varepsilon^H$. It can be shown that the sequential range of $B_H(t)$ will increase proportionally to N^H, where the sequential range is defined by

$$\text{sequential range} = \max_{t<r<t+N} B_H(r) - \min_{t<r<t+N} B_H(r) \qquad [10.4.1]$$

where t is continuous time and H is the model parameter. When the process $B_H(t)$ is Gaussian, it is called fractional Brownian motion. Discrete time *fractional Gaussian noise (FGN)* is defined for discrete time t by the increments

$$z_t = B_H(t + 1) - B_H(t) \qquad [10.4.2]$$

FGN is what Mandelbrot and Wallis (1969c) consider to be a model of Hurst's geophysical time series.

Mandelbrot and Van Ness (1968) and Mandelbrot and Wallis (1969a,b,c) have derived a number of properties of FGN. First, the parameter H must satisfy $0 < H < 1$. The sample mean and variance of FGN are consistent estimators of the true mean and variance, and FGN is covariance stationary. The expected values of the crude and adjusted ranges for FGN are the asymptotic relationships

$$E(R_N) = a_H N^H\,, \quad 0 < H < 1 \qquad [10.4.3]$$

and

$$E(R^*_N) = b_H N^H\,, \quad 0 < H < 1 \qquad [10.4.4]$$

where a_H and b_H are coefficients that do not depend on N. It can also be shown that for large N (Rao, 1973, p. 122),

$$E(\bar{R}^*_N) = aN^H \qquad [10.4.5]$$

Although the above asymptotic formulae are correct mathematically, they may possess limitations with respect to modelling Hurst's findings. Of foremost importance is the fact that Hurst examined $\bar{R}^*_N$ for small N and not the asymptotic expected values of R_N, R^*_N, and $\bar{R}^*_N$. Behaviour of any of the range statistics for large N does not necessarily infer the structure of $\bar{R}^*_N$ for small and moderate N. Even though [10.4.3] to [10.4.5] are asymptotically valid, in reality the Hurst coefficient is a function of N and is not a constant as is the parameter H in FGN. For example, as is shown by simulation experiments for NID random variables in Table 10.5.2, the expected value of the Hurst coefficient K is significantly larger than 0.5 for small N. A sequence of NID random variables is equivalent to a FGN process with $H = \frac{1}{2}$.

The theoretical ACF at lag k of FGN is given by

$$\rho_k = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}], \quad 0 < H < 1 \text{ and } k \geq 1 \qquad [10.4.6]$$

For large lags, [10.4.6] may be approximated by

$$\rho_k = H(2H-1)k^{2H-2} \qquad [10.4.7]$$

The $N \times N$ correlation matrix for FGN is given as

$$\mathbf{C}_N(H) = [\rho_{|i-j|}] \qquad [10.4.8]$$

where $\rho_0 = 1$ and ρ_k is calculated from [10.4.6] for $k \geq 1$. The Cholesky decomposition (Healy, 1968) of $\mathbf{C}_N(H)$ is determined such that

$$\mathbf{C}_N(H) = \mathbf{M}\mathbf{M}^T \qquad [10.4.9]$$

where $\mathbf{M}$ is the $N \times N$ lower triangular matrix having m_{ij} as a typical element. The matrix $\mathbf{M}$ is used for carrying out diagnostic checks, and simulating with FGN in Sections 10.4.4 and 10.4.6, respectively.

An examination of [10.4.6] and [10.4.7] reveals that $\rho_k \to 0$ as $k \to \infty$, but ρ_k is not summable if $\frac{1}{2} < H < 1$. Therefore, for $\frac{1}{2} < H < 1$, the FGN process has long memory. When $0 < H \leq \frac{1}{2}$, FGN constitutes a short memory process.

For many geophysical phenomena, the estimates for H are greater than $\frac{1}{2}$ but less than 1. Because the FGN model is not summable for H in this range, the statistical effect of past events on present behaviour attenuates very slowly. Therefore, long term persistence, as described by the theoretical ACF, is synonymous with $\frac{1}{2} < H < 1$. Some hydrologists claim that the form of the theoretical ACF for $\frac{1}{2} < H < 1$ is explained by the physical existence of an extremely long memory in hydrologic and other processes. But, as was pointed out by Klemes (1974), making inferences about physical features of a process based on operational models can be not only inaccurate but also misleading. Klemes correctly states that "... it must be remembered that the mathematical definition of FGN did not arise as a result of the physical or dynamic properties of geophysical and other processes but from a desire to describe an observed geometric pattern of historic time series mathematically... Thus FGN is an operational, not a physically founded model." Klemes demonstrates that the Hurst phenomenon could be due to zero memory nonstationary models and also specific types of storage systems. However, although physical interpretations that use operational models should be formulated and interpreted with caution, one criterion that is essential is that the statistical properties of any historical time series be incorporated properly into the stochastic model.

The appropriateness of long memory processes for modelling annual riverflow and other types of natural time series has been questioned previously by various hydrologists (Scheidegger, 1970; Klemes, 1974). Moreover, later in Section 10.4.7, it is shown that the FGN model can fail to provide an adequate statistical fit to historical annual riverflows.

The FGN model for a time series $z_t, t = 1,2,...,N$, can be specified in terms of the three parameters μ, γ_0, and H, where $E[z_t] = \mu$, $Var[z_t] = \gamma_0$, and the theoretical ACF of z_t is given by [10.4.6]. From these specifications, improved estimation and simulation procedures can be developed. Complete Fortran computer algorithms for these methods are given by Hipel and McLeod (1978b).

As explained in Part III in the book, when determining a long memory or a short memory model or in general any type of stochastic process for modelling a given data set, it is recommended to adhere to the three stages of model development. The first step consists of identifying, or choosing, the type of model to fit to the time series. If circumstances warrant the employment of a FGN process, then at the estimation stage, efficient MLE's of the model parameters can be procured by using the technique developed in Section 10.4.3. It is also shown in Section 10.4.4 how the model residuals of FGN can be calculated after the model parameters have been estimated. If diagnostic checks of the residuals reveal that modelling assumptions such as residual whiteness, normality, and homoscedasticity (i.e., constant variance) are not satisfied, then appropriate action can be taken. For example, a Box-Cox transformation (see Section 3.4.5) of the data prior to fitting a FGN process may rectify certain anomalies in the residuals. In some cases, a short memory model such as an ARMA process may provide a better statistical fit while at the same time preserve important historical statistics such as the RAR. The AIC (see Section 6.3) is recommended as a means of selecting the best model from a set of tentative models that may consist of both short memory and long memory processes.

10.4.3 Maximum Likelihood Estimation

In addition to the mean and variance, an estimate of the parameter H forms the only link that a FGN model has with the real world as represented by the historical data. Previously, various estimates for H have been formulated. Some researchers employ K in [10.3.4] as an estimate of H. Wallis and Matalas (1970) recommend the G Hurst statistics as an estimate of H. Unfortunately, little is known about the theoretical distribution of this estimate, and the G Hurst statistic in effect constitutes only an ad hoc method of calculating H. Young and Jettmar (1976, p. 830, equation 4) suggest a moment estimate for H based on an estimate of the historical ACF at lag 1 and [10.4.6]. They also develop a least squares estimate for H that is formulated by using the sample ACF and [10.4.6] (Young and Jettmar, 1976, p. 831, equation 6). However, McLeod and Hipel (1978b) question the theoretical basis and efficiency of Young and Jettmar's least squares estimate for H.

An alternative approach to estimating the parameters of a FGN model is to employ maximum likelihood estimation. The method of maximum likelihood procedure is widely used for the estimation of parametric models, since it often yields the most efficient estimates (see Section 6.2). Dunsmuir and Hannan (1976) show that the MLE's of the parameters of time series models often yield optimal estimates under very general conditions, which include the FGN model as a special case.

Given a historical time series $z_1, z_2, \ldots, z_N$, the log likelihood of μ, γ_0, and H in the FGN model is

$$\log L^*(\mu,\gamma_0,H) = -\frac{1}{2}\log|C_N(H)| - (2\gamma_0)^{-1}S(\mu,H) - (N/2)\log\gamma_0 \qquad [10.4.10]$$

where $C_N(H)$ is the correlation matrix given by [10.4.8]. The function $S(\mu,H)$ in [10.4.10] is determined by

$$S(\mu,H) = (\mathbf{z} - \mu\mathbf{1})^T[C_N(H)]^{-1}(\mathbf{z} - \mu\mathbf{1}) \qquad [10.4.11]$$

where $\mathbf{z}^T = (z_1,z_2,\ldots,z_N)$ is a $1\times N$ vector and $\mathbf{1}^T = (1,1,\ldots,1)$ is a $1\times N$ vector. For fixed H, the MLE of μ and γ_0 are

$$\hat{\mu} = \{\mathbf{z}^T[C_N(H)]^{-1}\mathbf{1}\}/\{\mathbf{1}^T[C_N(H)]^{-1}\} \qquad [10.4.12]$$

and

$$\hat{\gamma}_0 = N^{-1}S(\hat{\mu},H) \qquad [10.4.13]$$

Thus, the maximized log likelihood function of H is

$$\log L_{\max}(H) = -\frac{1}{2}\log|C_N(H)| - N/2\log[S(\hat{\mu},H)]/N \qquad [10.4.14]$$

The inverse quadratic interpolation search method can then be used to maximize $\log L_{\max}(H)$ to determine $\hat{H}$, the MLE of H. The variance of $\hat{H}$, given by $Var(\hat{H})$, is approximately

$$Var(\hat{H}) = -1\left|\frac{\partial^2\log L_{\max}(H)}{\partial H^2}\right|_{H=\hat{H}} \qquad [10.4.15]$$

The variance in [10.4.14] can be evaluated by numerical differentiation. If the computer algorithms given by Hipel and McLeod (1978b) are utilized, the computer time required for these calculations is not excessive provided that N is not too large (not larger than about 200). The standard error (SE) of the MLE of H is simply the square root of $Var(\hat{H})$ in [10.4.15].

In order to compare the statistical efficiency of the maximum likelihood and G Hurst estimation procedures, a simulation study is performed. For $H = 0.5$, 0.6, 0.7, 0.8, and 0.9 and for $N = 50$ and 100, 500 simulated series for each FGN model are generated by using the exact simulation technique given in Section 10.4.6. For each synthetic trace, the MLE $\hat{H}$ and the G Hurst estimate obtained by using $GH(10)$ as defined by Wallis and Matalas (1970) are determined. Because $GH(10)$ and $\hat{H}$ are functionally independent of the mean and variance, it is simplest to set the mean equal to zero and to assign the variance a value of unity when generating the synthetic data by using the method of Section 10.4.6. The *mean square errors (MSE)* of the maximum likelihood and $GH(10)$ estimators for a particular value of N are

$$MSE_{MLE}(H,N) = \frac{1}{500}\sum_{i=1}^{500}(\hat{H}_i - H)^2 \qquad [10.4.16]$$

and

$$MSE_{GH}(H,N) = \frac{1}{500}\sum_{i=1}^{500}(GH_i - H)^2 \qquad [10.4.17]$$

where $\hat{H}_i$ is the MLE of H for the ith simulated series of length N having a particular true value of H and GH_i is the magnitude of $GH(10)$ for the ith simulated series of length N with a specified true value of H.

The MSE criterion constitutes a practical overall measure for assessing the accuracy of an estimate. The MSE is equal to the square of the *bias* of the estimate plus the variance of the estimate. Because a biased estimator may in certain cases have smaller overall MSE, the "unbiasedness" of an estimate alone is not necessarily the most important requirement of an estimate. The relative efficiency (RE) of the $GH(10)$ estimate in comparison with the MLE $\hat{H}$ is

$$RE(H,N) = [MSE_{MLE}(H,N)]/[MSE_{GH}(H,N)] \qquad [10.4.18]$$

The entries in Table 10.4.1 confirm that the MLE procedure is significantly more accurate than the G Hurst method.

Table 10.4.1. Percentage relative efficiency of $GH(10)$ versus $\hat{H}$.

	N	
H	50	100
0.5	48	38
0.6	55	44
0.7	59	47
0.8	57	43
0.9	50	34

As explained in Section 6.3, the AIC is useful for discriminating among competing parametric models (Akaike, 1974). For the FGN model, the AIC is given by

$$AIC = -2\log L_{max}(H) + 4 \qquad [10.4.19]$$

When comparing models, the one with the smallest AIC provides the best statistical fit with the minimum number of model parameters.

10.4.4 Testing Model Adequacy

After fitting a statistical model to data, it is advisable to examine the chosen model for possible inadequacies which could seriously invalidate the model. The residuals of the FGN model with parameters μ, γ_0, and H can be defined by

$$\mathbf{e} = \mathbf{M}^{-1}(\mathbf{z} - \mu\mathbf{1}) \qquad [10.4.20]$$

where $\mathbf{e}^T = (e_1, e_2, \ldots, e_N)$ is the vector of model residuals. If the chosen model provides an adequate fit, the elements of $\mathbf{e}$ should be white noise that is NID(0,1). Accordingly, for any proposed FGN a suggested diagnostic check is to test the residuals for whiteness by employing suitable tests for whiteness (see Sections 7.3 and 2.6). For instance, the cumulative periodogram test of Section 2.6 could be utilized to check for residual whiteness. Other appropriate tests could be invoked to test whether the less important assumptions of normality (see Section 7.4) and homoscedasticity (see Section 7.5) of the residuals are also satisfied.

10.4.5 Forecasting with FGN

In Section 8.2, minimum mean square error (MMSE) forecasts are defined for use with ARMA models. One can, of course, also determine MMSE forecasts for FGN models. More specifically, Noakes et al. (1988) develop a formula for calculating one step ahead forecasts for a FGN model by employing the standard regression function (Anderson, 1958). First, to obtain the covariance matrix, Γ_N, one can substitute the MLE for H from Section 10.4.3 into [10.4.6] and then divide [10.4.8] by the estimated variance from [10.4.13]. The one step ahead forecast is then given by

$$E\{\mathbf{Z}_{N+1}|\mathbf{Z}_N\} = \mu + \gamma_N^T\Gamma_N^{-1}(\mathbf{Z}_N - \mu\mathbf{1}) \qquad [10.4.21]$$

where $\mathbf{Z}'_N = (z_1, z_2, \ldots, z_N)$, and $\gamma_N = (\gamma_N, \gamma_{N-1}, \ldots, \gamma_1)$. Rather than inverting Γ_N, let

$$\Gamma_N \mathbf{X}_N = (\mathbf{Z}_N - \mu\mathbf{1}) \qquad [10.4.22]$$

and solve for X_N. The solution of this system of equations is obtained using a Cholesky decomposition (Healy, 1968) of Γ_N such that

$$\mathbf{M}\mathbf{M}'\mathbf{X}_N = (\mathbf{Z}_N - \mu\mathbf{1}) \qquad [10.4.23]$$

where $\mathbf{M}$ is a $N \times N$ lower triangular matrix. The one step ahead forecast of Z_{N+1} is thus

$$E\{Z_{N+1}|\mathbf{Z}_N\} = \mu + \gamma_N \mathbf{X}_N \qquad [10.4.24]$$

Successive one step ahead forecasts can be obtained using the following procedure. Given $\mathbf{M}$, the covariance matrix for $\mathbf{Z}_{N+1}$ may be written as

$$\begin{aligned}\Gamma_{N+1} &= \begin{bmatrix}\Gamma_N & \gamma_N \\ \gamma_N^T & \gamma_0\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{M} & \mathbf{0} \\ \mathbf{a}_T & \alpha\end{bmatrix}\begin{bmatrix}\mathbf{M}' & \mathbf{a}' \\ \mathbf{0} & \alpha\end{bmatrix} \\ &= M^*M^{*\prime}\end{aligned} \qquad [10.4.25]$$

where $\mathbf{M}^*$ is a $(N+1)\times(N+1)$ lower triangular matrix. The Cholesky decomposition of Γ_{N+1} is calculated by noting that

$$\mathbf{M}\mathbf{a} = \gamma_N \qquad [10.4.26]$$

and

$$\alpha = \sqrt{\gamma_0 - \mathbf{a}^T\mathbf{a}}. \qquad [10.4.27]$$

Thus, the forecast of Z_{N+2} is given by

$$E\{Z_{N+2}|\mathbf{Z}_{N+1}\} = \mu + \gamma_{N+1}\mathbf{X}_{N+1} \qquad [10.4.28]$$

where $\mathbf{X}_{N+1}$ is obtained from

$$\mathbf{M}^*\mathbf{M}^{*T}\mathbf{X}_{N+1} = (\mathbf{Z}_{N+1} - \mu\mathbf{1}) \qquad [10.4.29]$$

10.4.6 Simulation of FGN

Historically, researchers have not developed an exact technique for simulating FGN. Instead, short memory approximations of FGN models have been utilized to generate synthetic traces. The methods used for obtaining approximate realizations of FGN include (1) type 1 (Mandelbrot and Wallis, 1969c), (2) type 2 (Mandelbrot and Wallis, 1969c), (3) fast FGN (Mandelbrot, 1971), (4) filtered FGN (Matalas and Wallis, 1971), (5) ARMA(1,1) (O'Connell, 1974a,b), (6) broken line (Rodriguez-Iturbe et al., 1972; Mejia et al., 1972; Garcia et al., 1972; Mandelbrot, 1972), and (7) ARMA-Markov (Lettenmaier and Burges, 1977) models.

Various papers have been written that include surveys and appraisals of one or more of the short memory approximations to FGN (see Lawrance and Kottegoda, 1977; Lettenmaier and Burges, 1977; O'Connell, 1974b; and Wallis and O'Connell, 1973). Although the underlying drawback of all these approximate processes is that the simulated data does not lie outside the Brownian domain (see Mandelbrot and Wallis (1968) for a definition of Brownian domain), additional handicaps of some of the models have also been cited in the literature. For instance, Lawrance and Kottegoda (1977) mention that the lack of a suitable estimation procedure for the parameters of a broken line process is the greatest deterrent to the utilization of that model by hydrologists.

When generating synthetic traces from a short memory approximation to FGN or any other type of stochastic model, proper simulation procedures should be adhered to. If more than one simulated time series from a certain model is needed, then it would be improper to first simulate one long synthetic time series and then to subdivide this longer trace into the required number of shorter time series. Rather, it would be more efficient to generate the shorter series independently so that the resulting estimates from each of the shorter series would be statistically independent. Furthermore, the standard errors of the particular parameters being estimated by the simulation study can be calculated if the estimates are statistically independent, but if they are correlated, the standard errors are not easily estimated. These and other guidelines for use in simulation are discussed in detail in Chapter 9.

Instead of the employment of short memory approximations for simulating FGN, it is possible to generate exact realizations of FGN. This procedure is analogous to the WASIM2 approach for simulating using ARMA models given in Section 9.4 and is based upon a knowledge of the theoretical ACF. Suppose that a FGN series $z_1, z_2, \ldots, z_N$, with parameters μ, γ_0, and H is to be simulated. Firstly, by utilizing an appropriate standard method, generate a Gaussian white noise sequence $e_1, e_2, \ldots, e_N$, that is NID(0,1) [see Section 9.2.3]. Next, calculate the $N \times N$ correlation matrix $\mathbf{C}_N(H)$, using [10.4.8]. Then, the Cholesky decomposition of $\mathbf{C}_N(H)$ is carried out to obtain the lower triangular matrix $\mathbf{M}$ in [10.4.9]. Exact realizations of FGN are calculated from

$$z_t = \mu + \left[\sum_{i=1}^{t} m_{ti} e_i \right] (\gamma_0)^{1/2} \qquad [10.4.30]$$

for $t = 1,2, \ldots, N$, and for $0 < H < 1$, where z_t is the FGN time series value that is $N(\mu,\gamma_0)$, and m_{ti} is from the matrix **M** in [10.4.9]. If the model parameter H is in the range $0.5 < H < 1$, then the synthesized data will lie outside the Brownian domain.

The computer algorithm for exactly simulating FGN is listed in standard Fortran by Hipel and McLeod (1978b). This algorithm requires only about $\frac{1}{2}N(N+2)$ storage locations to simulate a FGN series of length N. Thus, a modest requirement of about 5000 words is required to handle a series of length 100.

10.4.7 Applications to Annual Riverflows

Information concerning six of the longest annual riverflow time series given by Yevjevich (1963) is listed in Table 10.4.2. For each of these time series, the MLE of H in the FGN model and its SE are calculated. Table 10.4.3 lists the MLE and SE's (in parentheses) of H (see Section 10.4.3) and also the Hurst K (see [10.3.4]) and $GH(10)$ (Wallis and Matalas, 1970) estimates for each of the time series.

In Table 10.4.3, notice the difference between the three estimates of the FGN parameter H for each of the data sets. For instance, $\hat{H}$ for the Gota River has a magnitude of 0.839 with a corresponding SE of 0.073. Both the $GH(10)$ and K estimates for the Gota River are more than two times the SE less than the MLE of H.

The parameter estimates for the proper ARMA models that are fitted to the time series in Table 10.4.2 are given later in Table 10.6.3. Both the Danube River and the Rhine River time series are simply white noise. If a time series is NID, the theoretical value of H for a FGN model is 0.5. For both the Danube River and the Rhine River, Table 10.4.3 reveals that the MLE of H is closer to 0.5 than either the $GH(10)$ or the K estimate. In addition, for each of the two data sets, $\hat{H}$ is easily within one SE of 0.5.

In order to determine whether a short memory or a long memory model should be selected for each of the six time series, the AIC can be utilized (see Section 6.3). Table 10.4.4 lists the values of the AIC for the FGN models by using $\hat{H}$ and the best fitting ARMA model. For each of the six cases, the AIC for the ARMA model has a magnitude less than that for the FGN model. Therefore, on the basis of a combination of best statistical fit and model parsimony, the ARMA model should be chosen in preference to the FGN process for the time series considered.

The Gota River is instructive for portraying possible problems that may arise when using FGN models in practice, since it appears that no FGN model can give an adequate fit to this time series. After a FGN model has been fit to a given data set, it is recommended to implement appropriate diagnostic checks for testing model adequacy. It is of utmost importance that the residuals of FGN given by [10.4.20] be white noise. Accordingly, plots of the cumulative periodogram from Section 2.6 for the residuals of the FGN models for the Gota River obtained by using $\hat{H}$, $GH(10)$, and K are displayed in Figures 10.4.1 to 10.4.3, respectively. The 1%, 5%, 10% and 25% significance levels are indicated on the plots. As is shown in the figures, the cumulative periodogram test is significant in all three cases at the 1% level, although the departure from whiteness is not as great for the FGN model when using $\hat{H}$ as it is for the other two

Table 10.4.2. Average annual riverflow time series.

Code Name	River	Location	Period	N
Mstouis	Mississippi	St. Louis, Missouri	1861-1957	96
Neumunas	Neumunas	Smalinikai, U.S.S.R.	1811-1943	132
Danube	Danube	Orshava, Romania	1837-1957	120
Rhine	Rhine	Basle, Switzerland	1807-1957	150
Ogden	St. Lawrence	Ogdensburg, New York	1860-1957	97
Gota	Gota	Sjotorp-Vanersburg, Sweden	1807-1957	150

Table 10.4.3. Estimated statistics for the annual riverflows.

Data Set	$\hat{H}$	$GH(10)$	K
Mstouis	0.674 (0.082)	0.580	0.648
Neumunas	0.591 (0.067)	0.520	0.660
Danube	0.548 (0.063)	0.560	0.633
Rhine	0.510 (0.058)	0.592	0.614
Ogden	0.949 (0.047)	0.868	0.894
Gota	0.839 (0.073)	0.523	0.689

*The parenthetical values are SE's.

Table 10.4.4. AIC values for the fitted FGN and ARMA models.

Data Set	FGN Models	ARMA Models
Mstouis	1400.0	1395.8
Neumunas	1207.5	1198.2
Danube	1666.7	1389.0
Rhine	1531.8	1529.8
Ogden	1176.9	1172.1
Gota	1350.6	1331.0

cases. Therefore, the whiteness diagnostic checks indicate that because of the dependence of the model residuals the FGN processes provide a poor statistical fit to the given data. Hence, it would be advisable to consider another type of process to model the annual riverflows of the Gota River.

When selecting a process to describe a given time series, it is highly desirable that important historical statistics such as the ACF at various lags (especially at low lags for nonseasonal models) be preserved. The inability of the three FGN models for the Gota River to pass the diagnostic check for residual whiteness precludes the preservation of historical statistics by these models. The sample ACF of the Gota River is shown in Figure 10.4.4, while the theoretical ACF of FGN, obtained by using $\hat{H}$ and K, are displayed in Figures 10.4.5 and 10.4.6, respectively. To calculate the theoretical ACF for FGN in Figures 10.4.5 and 10.4.6, the values of $\hat{H}$ and K for the Gota River in Table 10.4.3 are substituted into [10.4.6]. Because the theoretical ACF of FGN obtained by using the $GH(10)$ estimate is not significantly different from 0.5, the plot of this theoretical ACF would be very close to white noise and is, therefore, not given. Nevertheless, comparisons of Figure 10.4.4 with Figures 10.4.5 and 10.4.6 reveal visually that the historical sample ACF is not preserved by the FGN models.

In contrast to the inability of a FGN process to model the Gota riverflows, an ARMA model does provide an adequate fit to the data. By following the identification, estimation, and diagnostic check stages of model construction presented in Part III, the best type of ARMA model to describe the Gota riverflows is an ARMA process with two AR parameters (denoted by ARMA(2,0)). The ARMA(2,0) process provides a slightly better fit than an ARMA model with one MA parameter (denoted as ARMA(0,1)). The AIC also selects the ARMA(2,0) model in preference to the ARMA(0,1) model. In addition, the ARMA(2,0) model passes rigorous diagnostic checks for whiteness, homoscedasticity, and normality of the model residuals.

By knowing the parameter estimates of an ARMA model, it is possible to calculate the theoretical ACF by employing a technique described in Appendix A3.2. Figure 10.4.7 is a plot of the theoretical ACF for the ARMA(2,0) model for the Gota River data. A comparison of Figures 10.4.7 and 10.4.4 demonstrates that the ARMA model preserves the historical ACF especially at the important lower lags. Notice that the value of the ACF for lags 1 to 4 are almost identical for these two plots.

In addition to the use of graphical aids to determine whether historical statistics are preserved, a more rigorous procedure can be followed. In Section 10.6 a statistical test is used in conjunction with Monte Carlo techniques in order to determine the ability of a class of models to preserve specified historical statistics. It is demonstrated that ARMA processes preserve the RAR or equivalently K. This procedure could also be adopted for statistics such as various lags of the ACF to show quantitatively whether or not these statistics are preserved by the calibrated models.

The inability of a FGN process to preserve the ACF and perhaps other historical statistics in some practical applications could be due to the inherent mathematical structure and underlying properties that were discussed previously. Another obvious drawback of FGN is the dependence of the model on only a few parameters. In addition to the mean and variance, an estimate of the parameter H forms the only actual link between the theoretical model and the real world as presented by the data. This renders FGN processes highly inflexible. On the other hand, in ARMA modelling the form of the model is tailored specifically to fit a given set of data. At the

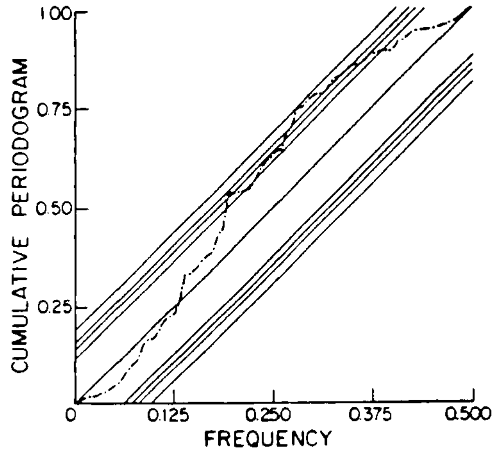


Figure 10.4.1. Gota River residual cumulative periodogram for the FGN model using $\hat{H}$.

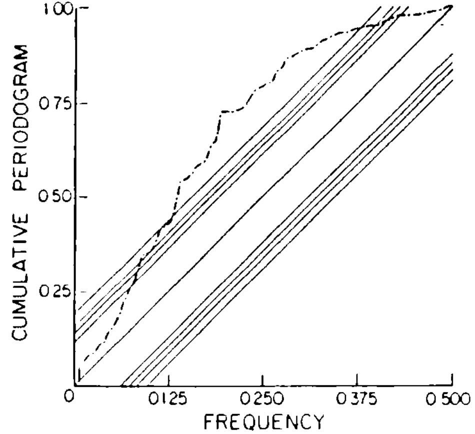


Figure 10.4.2. Gota River residual cumulative periodogram for the FGN model using $GH(10)$.

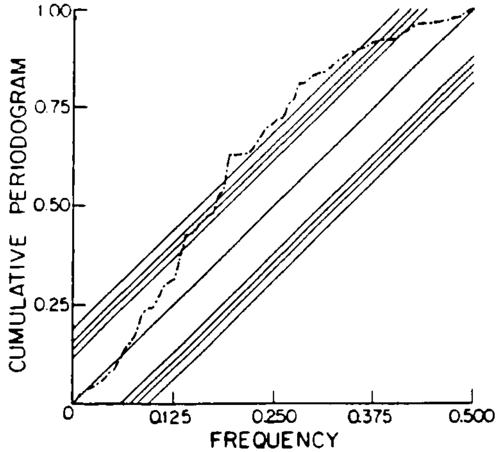


Figure 10.4.3. Gota River residual cumulative periodogram for the FGN model using K.

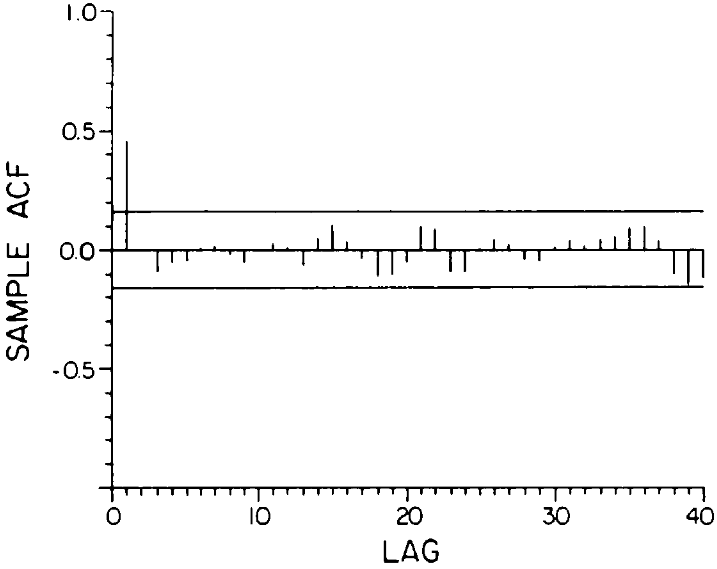


Figure 10.4.4. Sample ACF of the Gota River.

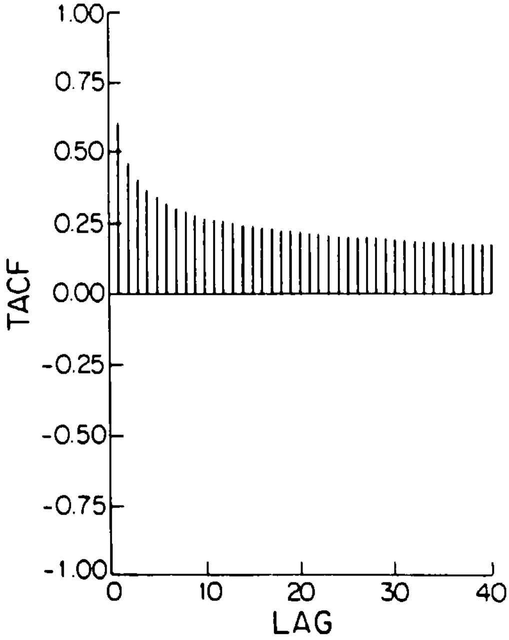


Figure 10.4.5. Theoretical ACF of the Gota River for the FGN model using $\hat{H}$.

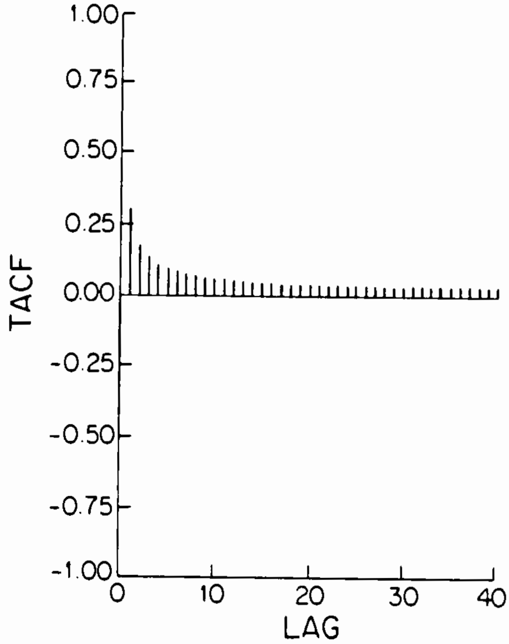


Figure 10.4.6. Theoretical ACF of the Gota River for the FGN model using K.

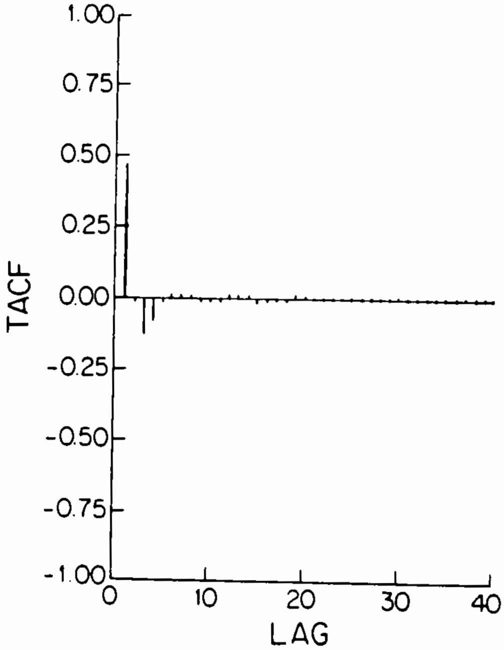


Figure 10.4.7. Theoretical ACF of the Gota River for the ARMA(2,0) model.

identification stage, the general structure of the data is determined by observing the shape of the ACF and other graphs described in Section 5.3. An appropriate number of AR and MA parameters are selected in order that the selected ARMA model fits the data as closely as possible using a minimum number of parameters. Rigorous checks are performed to insure that the white noise component of the model is not correlated. If all the modelling assumptions are satisfied, this guarantees that important historical statistics such as the ACF, the RAR, and K will be preserved reasonably well by the model.

10.5 SIMULATION STUDIES

10.5.1 Introduction

When studying statistics such as the RAR and K, information is required regarding first, second, and perhaps higher order moments of the statistics. In general, it would be most advantageous to know the exact distribution of the statistic under study. Three approaches are available to obtain knowledge regarding the mathematical properties of a specified statistic. One method is to *derive an exact analytical expression* for the moments and perhaps the distribution of the statistic. Except for special cases of the lower order moments of a statistic, this precise procedure is often analytically intractable. Only after extensive research, Anis and Lloyd (1976) were able to derive in [10.3.15] the exact expression for the expected value of the RAR for NID summands.

A second approach is to *develop asymptotic formulae* for the distributional properties of a given statistic. This approximate procedure may yield results that are useful in certain situations, while in other circumstances the output may suffer from lack of accuracy, especially for small N. Feller (1951), for example, proved an asymptotic relationship that is valid for the expected value of the adjusted range and also the RAR of IID random variables (see [10.3.5] and [10.3.6], respectively). Siddiqui (1976) derived asymptotic expressions for calculating the expected value of the RAR for any short memory process.

In the third approach, *simulation* is used to determine as accurately as desired the distributional attributes of a given statistic. In Section 10.6, Monte Carlo procedures are utilized to obtain the empirical distribution of the RAR and K. Although some researchers may argue that simulation may be relatively costly with respect to computer usage, the fact of the matter is that answers are needed now to help solve present day engineering problems. In addition, because of the vast mathematical complexity that is often required to prove exact analytical solutions, simulation results may help to economize academic endeavours by delineating the more promising avenues of research that could also be scrutinized analytically. Finally, it should be borne in mind that in comparison with an exact analytical solution, simulation provides a straightforward but equally correct resolution to the problem of the distributional characteristics of a particular statistic. The theory and practice of simulating with ARMA models are discussed in detail in Chapter 9 while an exact simulation method for use with FGN is given in this chapter in Section 10.4.6.

The simulation investigations of this section deal primarily with the estimated mean and variance of a certain statistic. Suppose that $\overline{N}$ independent simulations of a time series $z_1, z_2, \ldots, z_N$, are obtained and that a statistic $T = T(z_1, z_2, \ldots, z_N)$ is calculated in each simulated series. The empirical *mean* of T is then given by

$$\bar{T} = \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} T_i \qquad [10.5.1]$$

where T_i is the value of T in the ith simulation. If each successive realization of the sequence $z_1, z_2, \ldots, z_N$, is independent of previous realizations so that the T_i are statistically independent, then the *variance* of T can be estimated by

$$V_T = \frac{1}{\bar{N}-1}\sum_{i=1}^{\bar{N}} (T_i - \bar{T})^2 \qquad [10.5.2]$$

By the central limit theorem, $\bar{T}$ will be distributed very nearly normally with mean equal to $E(\bar{T})$ and with variance approximately equal to $V_T/\bar{N}$. Thus, the standard deviation and confidence intervals of the expected value being estimated are readily obtained.

If $\bar{N}$ white noise series of length N are being simulated, then it is correct to simulate a single time series of length $\bar{N}N$ and then subdivide it into $\bar{N}$ series with N values in each series. However, if a correlated series is being simulated, the aforementioned procedure should not be followed. For instance, if $\bar{N}$ FGN series with $0.5 < H < 1$ are being formulated by first generating a long series of length $\bar{N}N$ and then subdividing this into $\bar{N}$ subsequences of length N, then the resulting T_i will in general be correlated. Therefore, the resulting estimate for $E(T)$ in [10.5.1] will be less precise (i.e., have larger variance), and the estimate of the variance of T in [10.5.2] will be underestimated, so that correct standard deviations and confidence intervals for $E(T)$ will not be available.

10.5.2 Simulation of Independent Summands

The Rescaled Adjusted Range

Mandelbrot and Wallis (1969e) reported simulation experiments which indicated that the expected value of the RAR for IID summands is virtually independent of the underlying distribution. However, as was pointed out by Taqqu (1970), the simulation study of Mandelbrot and Wallis (1969e) contained a serious programming error in the calculation of the RAR. Accordingly, another study of the robustness of the expected value of the RAR with respect to the underlying distribution is required.

A simulation study is performed for various types of white noise series varying in length from $N = 5$ to $N = 200$. For each value of N, the number of series of length N that are generated is $\bar{N} = 10{,}000$. The expected values of the RAR are determined by using [10.5.1] for the following independent summands: (1) normal, (2) gamma with shape parameter 0.1, (3) symmetric stable with characteristic exponent $\alpha = 1.3$, and (4) Cauchy. The simulation results for $E(\bar{R}^*_N)$ at specific values of N for the aforementioned summands are listed in Table 10.5.1. The standard deviations of the estimated values of $E(\bar{R}^*_N)$ are determined by using the square root of [10.5.2] and are given in parentheses below the estimates in Table 10.5.1. The exact values of $E(\bar{R}^*_N)$ for NID random variables are calculated by using the formula of Anis and Lloyd (1976) that is written in [10.3.15]. Comparisons of columns 2 and 4 to 7 reveal that the expected value of the RAR is indeed rather insensitive to the underlying distribution for the values of N that are considered. Even for Cauchy summands, the expected value and variance of the RAR are quite similar to the NID case. The asymptotic results of Feller (1951) for $E(\bar{R}^*_N)$ of IID summands are determined by using [10.3.6] and are tabulated in Table 10.5.1. A perusal of the asymptotic and other entries in the table discloses that the approximation given by Feller's results improves with increasing N.

Anis and Lloyd (1975) developed analytical formulae for the expected value of the crude and adjusted ranges of independent gamma random variables. For highly skewed independent gamma summands, the local Hurst coefficient for the crude and adjusted ranges possessed values greater than 0.5 for N less than 1000. However, the results of Table 10.5.1 indicate that the expected values of the RAR for IID summands are approximately independent of the underlying distribution even if that distribution is gamma. Therefore, as was confirmed by O'Connell (1976), Anis and Lloyd's (1975) results do not hold for the RAR. In addition, Hurst studied K for the RAR and not the Anis and Lloyd local Hurst coefficient for the crude and adjusted ranges.

The Hurst Coefficient

As was mentioned previously, the Hurst statistic of primary import is the RAR. Nevertheless, because the Hurst coefficient K has been extensively investigated during the past quarter of a century, this fact may insure the survival of K as an important hydrological statistic for some time to come. Therefore, some statistical properties of K and other exponents are investigated.

First, it should be noted that because of the research results of Anis and Lloyd (1976) in [10.3.15], K can be evaluated analytically for NID summands. Let K' be the Hurst coefficient calculated by using

$$K' = \log E(\bar{R}^*_N)/(\log N - \log 2) \qquad [10.5.3]$$

where $E(\bar{R}^*_N)$ is determined exactly by using [10.3.15]. It follows from Jensen's inequality (Rao, 1973, p. 57) that for finite N,

$$E(K) < K' \qquad [10.5.4]$$

In Table 10.5.2, the magnitudes of K' from [10.5.3] are listed for the length of series N ranging from 5 to 200. When 10,000 series are generated for NID random variables for each N, then the expected value of K can be estimated by utilizing [10.5.1], while the standard deviation of $E(K)$ can be calculated by using the square root of [10.5.2]. In Table 10.5.2, the estimated values of $E(K)$ for various time series lengths are catalogued. The standard deviation of each estimate is contained in the parentheses below the estimate. A comparison of columns 2 and 3 in Table 10.5.2 demonstrates that the inequality in [10.5.4] is valid. However, the difference between $E(K)$ and K' is negligible. Therefore, [10.5.3] provides a viable means for estimating the expected value of K for NID summands. In addition, the Hurst coefficient K is obviously a function of the sample size, and for increasing N the coefficient K attenuates toward its asymptotic value of 0.5. However, for small and moderate values of N, the statistic K is significantly larger than 0.5.

The coefficient K constitutes one method of estimating the generalized Hurst coefficient h in [10.3.2]. Another approach is to evaluate h by using the estimate YH of Gomide (1975) that is given in [10.3.7]. By taking logarithms of [10.3.7], an explicit expression for YH is

Table 10.5.1. Expected values of the RAR for some IID summands.

N	Analytical Results		Simulation Results*			
	Anis and Lloyd (1976)	Feller (1951)	Normal	Gamma	Stable	Cauchy
5	1.9274	2.8025	1.9273	1.9851	1.9264	1.9506
			(0.0027)	(0.0018)	(0.0022)	(0.0026)
10	3.0233	3.9633	3.0302	3.0330	2.9699	3.0556
			(0.0060)	(0.0039)	(0.0047)	(0.0056)
15	3.8812	4.8541	3.8826	3.8356	3.7571	3.8987
			(0.0084)	(0.0056)	(0.0064)	(0.0079)
20	4.6111	5.6050	4.6047	4.5141	4.4408	4.6214
			(0.0100)	(0.0071)	(0.0075)	(0.0098)
25	5.2576	6.2666	5.2540	5.1213	5.0044	5.2889
			(0.0116)	(0.0085)	(0.0088)	(0.0115)
30	5.8443	6.8647	5.8770	5.6709	5.5681	5.8767
			(0.0131)	(0.0097)	(0.0098)	(0.0130)
35	6.3851	7.4147	6.4214	6.1707	6.0090	6.3974
			(0.0145)	(0.0109)	(0.0106)	(0.0143)
40	6.8895	7.9267	6.8920	6.6605	6.5037	6.9075
			(0.0158)	(0.0121)	(0.0118)	(0.0155)
45	7.3640	8.4075	7.3595	7.0938	6.9010	7.3934
			(0.0169)	(0.0132)	(0.0125)	(0.0166)
50	7.8133	8.8623	7.7785	7.5012	7.3184	7.8540
			(0.0180)	(0.0141)	(0.0132)	(0.0178)
60	8.6502	9.7081	8.6246	8.3061	8.0670	8.6263
			(0.0198)	(0.0159)	(0.0148)	(0.0195)
70	9.4210	10.4860	9.4453	9.0632	8.7242	9.4454
			(0.0215)	(0.0178)	(0.0158)	(0.0211)
80	10.1392	11.2100	10.1349	9.7327	9.3732	10.1336
			(0.0233)	(0.0194)	(0.0172)	(0.0232)
90	10.8143	11.8900	10.8208	10.4068	9.9544	10.8857
			(0.0248)	(0.0209)	(0.0183)	(0.0248)
100	11.4533	12.5331	11.4775	10.9769	10.5593	11.4546
			(0.0262)	(0.0224)	(0.0196)	(0.0258)
125	12.9243	14.0125	12.9617	12.4280	11.8353	12.9619
			(0.0299)	(0.0255)	(0.0220)	(0.0292)
150	14.2556	15.3499	14.1956	13.6864	13.0622	14.2636
			(0.0323)	(0.0285)	(0.0240)	(0.0323)
175	15.4806	16.5798	15.4198	14.8752	14.1069	15.4971
			(0.0349)	(0.0315)	(0.0261)	(0.0354)
200	16.6214	17.7245	16.5938	15.9992	15.1381	16.6259
			(0.0376)	(0.0337)	(0.0281)	(0.0376)

*The parenthetical values are standard deviations

Table 10.5.2. Hurst coefficients for NID summands.

N	K'	$E(K)$*	YH'
5	0.7161	0.7032 (0.0016)	0.3375
10	0.6874	0.6750 (0.0013)	0.4315
15	0.6731	0.6629 (0.0011)	0.4591
20	0.6638	0.6540 (0.0010)	0.4725
25	0.6571	0.6469 (0.0009)	0.4805
30	0.6519	0.6420 (0.0008)	0.4859
35	0.6477	0.6385 (0.0008)	0.4897
40	0.6442	0.6365 (0.0007)	0.4926
45	0.6413	0.6335 (0.0007)	0.4948
50	0.6387	0.6305 (0.0007)	0.4967
60	0.6344	0.6270 (0.0007)	0.4994
70	0.6309	0.6235 (0.0006)	0.5014
80	0.6279	0.6213 (0.0006)	0.5029
90	0.6254	0.6186 (0.0006)	0.5040
100	0.6233	0.5156 (0.0006)	0.5049
125	0.6189	0.6129 (0.0005)	0.5066
150	0.6154	0.6100 (0.0005)	0.5078
175	0.6127	0.6070 (0.0005)	0.5086
200	0.6103	0.6051 (0.0005)	0.5092

*The parenthetical values are standard deviations.

$$YH = (\log\bar{R}^*_N - \log 1.2533)/\log N \qquad [10.5.5]$$

Although the expected value of YH could be determined from simulation experiments, an alternative analytical procedure is to substitute $E(\bar{R}^*_N)$ from [10.3.15] for $\bar{R}^*_N$ in [10.5.5] and then to estimate YH by YH' by using [10.5.5]. In Table 10.5.2, the values of YH' are tabulated for different time series lengths. It is obvious that YH' is a function of the sample size and that YH' provides a closer approximation to the limiting value of 0.5 than does K.

10.5.3 Simulation of Correlated Summands

Long Memory Models

By utilizing [10.4.30], it is possible to simulate exactly synthetic traces of FGN. Because only short memory approximations to FGN processes were previously available for simulation purposes, the exact method should prove useful for checking former FGN simulation studies and also for exploring new avenues of research for long memory models. Of particular importance are Monte Carlo studies to investigate the statistical properties of FGN processes. Consider, for example, the behaviour of the RAR for FGN models. For time series varying in length from $N = 5$ to $N = 200$ a total of 10,000 simulated sequences are generated for each value of N. Because the RAR statistic is not a function of the mean and variance of a FGN process, it is convenient to assign the mean a value of zero and the variance a magnitude of 1 when performing the simulations using [10.4.30]. By utilizing [10.5.1] and [10.5.2], the expected values of the RAR and variances, respectively, are calculated. Table 10.5.3 records the estimates of $E(\bar{R}^*_N)$ and the corresponding standard deviations in brackets for FGN models with $H = 0.7$ and 0.9. From an inspection of the entries in Table 10.5.3, it is obvious that $E(\bar{R}^*_N)$ increases in magnitude for larger N. Furthermore, at a given value of N the expected value of the RAR is greater for a FGN model with $H = 0.9$ than it is a for a FGN process with $H = 0.7$.

Short Memory Models

In Chapter 9, improved procedures are given for generating synthetic traces using ARMA models. In particular, the WASIM1 (see Section 9.3) and WASIM2 (Section 9.4) procedures are recommended for use with ARMA models. When either WASIM1 or WASIM2 is employed, random realizations of the process under consideration are used as starting values. Since fixed initial values are not utilized, systematic bias is avoided in the generated data.

As a typical example of a short memory process, consider the Markov model of Section 3.2.1 given by

$$z_t = \phi_1 z_{t-1} + a_t \qquad [10.5.6]$$

where t equals $1, 2, \ldots, N$, ϕ_1 is the AR parameter, and a_t is the white noise that is NID$(0,\sigma_a^2)$. By using WASIM2, a total of 10,000 synthetic sequences are generated for specific values of N for Markov processes with $\phi_1 = 0.3$, 0.5 and 0.7. Because the RAR is independent of the variance of the innovations, a value such as unity may be used for σ_a^2 in the simulation study. In Tables 10.5.4 to 10.5.6, the expected values of the RAR and corresponding standard deviations in parentheses are given for the three Markov models. Comparisons of the third columns in these tables reveal that the expected value of the RAR increases for increasing N and ϕ_1.

Table 10.5.3. Expected values of the RAR for FGN models.

N	FGN Models*	
	H=0.7	*H*=0.9
5	1.9682	2.0100
	(0.0026)	(0.0025)
10	3.2716	3.5031
	(0.0062)	(0.0061)
15	4.3946	4.8751
	(0.0091)	(0.0094)
20	5.3972	6.1579
	(0.0116)	(0.0125)
25	6.3351	7.4051
	(0.0141)	(0.0155)
30	7.2066	8.6032
	(0.0165)	(0.0187)
35	8.0515	9.7839
	(0.0188)	(0.0216)
40	8.8767	10.9431
	(0.0205)	(0.0241)
45	9.6650	12.0926
	(0.0227)	(0.0271)
50	10.4007	13.2284
	(0.0247)	(0.0298)
60	11.8233	15.3575
	(0.0280)	(0.0352)
70	13.2003	17.4965
	(0.0322)	(0.0413)
80	14.5205	19.5945
	(0.0356)	(0.0461)
90	15.7709	21.6075
	(0.0389)	(0.0518)
100	16.9241	23.5818
	(0.0420)	(0.0573)
125	19.8877	28.5197
	(0.0494)	(0.0700)
150	22.6178	33.2646
	(0.0571)	(0.0831)
175	25.2291	38.0410
	(0.0638)	(0.0964)
200	27.7601	42.6710
	(0.0701)	(0.1080)

*The parenthetical values are standard deviations.

Table 10.5.4. Expected values of the RAR for a Markov model with $\phi_1 = 0.3$.

N	$E(\bar{R}^*_N)$	
	Asymptotic	Simulated*
5	3.8192	1.9875 (0.0026)
10	5.4011	3.3410 (0.0062)
15	6.6150	4.4633 (0.0089)
20	7.6383	5.4261 (0.0114)
25	8.5390	6.2853 (0.0135)
30	9.3550	7.0666 (0.0156)
35	10.1045	7.7976 (0.0175)
40	10.8022	8.5022 (0.0188)
45	11.4575	9.1493 (0.0205)
50	12.0772	9.7347 (0.0221)
60	13.2299	10.8709 (0.0242)
70	14.2900	11.9207 (0.0273)
80	15.2766	12.9177 (0.0296)
90	16.2033	13.8181 (0.0317)
100	17.0798	14.6243 (0.0335)
125	19.0958	16.6970 (0.0380)
150	20.9184	18.5288 (0.0424)
175	22.5944	20.1758 (0.0459)
200	24.1545	21.7339 (0.0491)

*The parenthetical values are standard deviations.

Table 10.5.5. Expected values of the RAR for a Markov model with $\phi_1 = 0.5$.

N	$E(\bar{R}^*_N)$	
	Asymptotic	Simulated*
5	4.8541	2.0194 (0.0025)
10	6.8647	3.5438 (0.0061)
15	8.4075	4.8738 (0.0092)
20	9.7081	6.0432 (0.0120)
25	10.8540	7.1131 (0.0147)
30	11.8900	8.0779 (0.0171)
35	12.8426	8.9858 (0.0194)
40	13.7294	9.8655 (0.0212)
45	14.5622	10.6837 (0.0233)
50	15.3499	11.4170 (0.0252)
60	16.8150	12.8455 (0.0283)
70	18.1622	14.1721 (0.0320)
80	19.4163	15.4320 (0.0350)
90	20.5941	16.5726 (0.0376)
100	21.7080	17.5991 (0.0400)
125	24.2703	20.2124 (0.0459)
150	26.5868	22.5342 (0.0515)
175	28.7170	24.6356 (0.0562)
200	30.6998	26.6039 (0.0603)

*The parenthetical values are standard deviations.

Table 10.5.6. Expected values of the RAR for a Markov model with $\phi_1 = 0.7$.

N	$E(\bar{R}^*_N)$	
	Asymptotic	Simulated*
5	6.6713	2.0435 (0.0025)
10	9.4346	3.7235 (0.0059)
15	11.5550	5.2915 (0.0091)
20	13.3425	6.7304 (0.0123)
25	14.9174	8.0874 (0.0154)
30	16.3412	9.3309 (0.0184)
35	17.6505	10.5117 (0.0212)
40	18.8692	11.6603 (0.0235)
45	20.0138	12.7462 (0.0262)
50	21.0964	13.7239 (0.0286)
60	23.1100	15.6339 (0.0331)
70	24.9616	17.4191 (0.0378)
80	26.6851	19.1225 (0.0419)
90	28.3038	20.6666 (0.0454)
100	29.8348	22.0685 (0.0490)
125	33.3564	25.6001 (0.0570)
150	36.5401	28.7578 (0.0648)
175	39.4678	31.6509 (0.0716)
200	42.1928	34.3412 (0.0772)

*The parenthetical values are standard deviations.

It is also possible to compare the estimated expected value of the RAR for a Markov model to an analytical large-sample approximation that is given by Siddiqui (1976) as

$$E(\bar{R}^*_N) = \{(\pi N/2)[(1-\phi_1^2)/(1-\phi_1)^2]\}^{1/2} \qquad [10.5.7]$$

In Tables 10.5.4 to 10.5.6, the output from [10.5.7] for the three Markov models are catalogued. A perusal of these tables demonstrates that Siddiqui's approximation for $E(\bar{R}^*_N)$ is not too accurate for the cases considered, and the precision decreases for increasing ϕ_1.

10.6 PRESERVATION OF THE RESCALED ADJUSTED RANGE

10.6.1 Introduction

A major challenge in stochastic hydrology is to determine models that preserve important historical statistics such as the rescaled adjusted range (RAR), or equivalently the Hurst coefficient K. The major finding of this section is that ARMA models do statistically preserve the historical RAR statistics or equivalently the Hurst coefficients denoted using K's. This interesting scientific result is what solves the riddle of the Hurst Phenomenon.

After fitting ARMA models to 23 annual geophysical time series, simulation studies are carried out to determine the small sample *empirical cumulative distribution function (ECDF)* of the RAR or K for various ARMA models. The ECDF for each of these statistics is shown to be a function of the time series length N and the parameter values of the specific ARMA process being considered. Furthermore, it is possible to determine as accurately as desired the distribution of the RAR or K. A theorem is given to obtain confidence intervals for the ECDF in order to guarantee a prescribed precision. Then it is shown by utilizing simulation results and a given statistical test that ARMA models do preserve the observed RAR or K of the 23 geophysical time series.

10.6.2 ARMA Modelling of Geophysical Phenomena

In this section, ARMA models are determined for 23 yearly geophysical time series. Table 10.6.1 lists the average annual riverflows and miscellaneous geophysical phenomena that are modelled. The riverflows are the longer records that are available in a paper by Yevjevich (1963). Although the flows were converted to cubic meters per second, it is irrelevant which units of measurement are used, since the AR and MA parameter estimates for the ARMA models fitted to the data are independent of the measuring system used. The mud varve, temperature, rainfall, sunspot numbers, and minimum flows of the Nile River are obtained from articles by De Geer (1940), Manley (1953, pp. 255-260), Kendall and Stuart (1963, p. 343), Waldmeier (1961), and Toussoun (1925), respectively.

Table 10.6.2 lists 12 sets of tree ring indices comprising six different species of trees from western North America. The indices labelled Snake are from a book by Schulman's (1956, p. 77), and the rest were selected from a report by Stokes et al. (1973).

By employing the three stages of model construction presented in detail in Part III, the most appropriate ARMA model from [3.4.32] is fitted to each of the 23 time series. Table 10.6.3 catalogues the type of ARMA model, Box-Cox transformation from [3.4.30], parameter estimates and standard errors (SE's) for each data set. The SE's are given in parentheses. For all the Box-Cox transformations, the constant is set equal to zero. When $\lambda = 1$ there is no

Table 10.6.1. Annual riverflows and miscellaneous geophysical data.

Code Name	Type	Location	Period	Length N
Mstouis	Mississippi River	St. Louis, Missouri	1861-1957	96
Neumunas	Neumunas River	Smalininkai, USSR	1811-1943	132
Danube	Danube River	Orshava, Romania	1837-1957	120
Rhine	Rhine River	Basle, Switzerland	1807-1957	150
Ogden	St. Lawrence River	Ogdensburg, New York	1860-1957	97
Gota	Gota River	Sjotorp-Vanersburg, Sweden	1807-1957	150
Espanola	mud varves	Espanola, Ontario	-471 to -820	350
Temp	temperature data	English Midlands	(Swedish time)	
Precip	precipitation	London, England	1698-1952	255
Sunyr	yearly sunspots	sun	1813-1912	100
Minimum	minimum flows of	Rhoda, Egypt	1798-1970	163
	Nile River		622-1469	848

Table 10.6.2. Tree ring indicies data.

Code Name	Type of Tree	Location	Period	Length N
Snake	Douglas fir	Snake River Basin	1282-1950	669
Exshaw	Douglas fir	Exshaw, Alberta, Canada	1460-1965	506
Naramata	Ponderosa pine	Naramata, B.C., Canada	1451-1965	515
Dell	Limber pine	Dell, Montana	1311-1965	655
Lakeview	Ponderosa pine	Lakeview, Oregon	1421-1964	544
Ninemile	Douglas fir	Nine Mile Canyon, Utah	1194-1964	771
Eaglecol	Douglas fir	Eagle, Colorado	1107-1964	858
Navajo	Douglas fir	Navajo National Monument (Belatakin), Arizona	1263-1962	700
Bryce	Ponderosa pine	Bryce Water Canyon, Utah	1340-1964	625
Tioga	Jeffrey pine	Tioga Pass, California	1304-1964	661
Bigcone	Big cone spruce	Southern California	1458-1966	509
Whitemtn	Bristlecone pine	White Mountains, California	800-1963	1164

transformation, while $\lambda = 0$ means that natural logarithms are taken of the data. Whenever a MLE of λ is calculated, the SE is included in parentheses.

10.6.3 Distribution of the RAR or *K*

Suppose the determination of the exact distribution of the RAR (i.e., $\bar{R}^*_N$) or K is required. The expected value of $\bar{R}^*_N$ is now known theoretically for both an independent and a symmetrically correlated Gaussian process (Anis and Lloyd, 1976). At present, the cumulative

Table 10.6.3. ARMA models fitted to the geophysical data.

Code Name	Model	λ*	Parameter	Value*	Parameter	Value*	Parameter	Value*
Mstouis	(0,1)	1.0	θ_1	-0.309				
				(0.094)				
Neumunas	(0,1)	0.0	θ_1	-0.222				
				(0.086)				
Danube	(0,0)	1.0						
Rhine	(0,0)	1.0						
Ogden	(3,0)	1.0	ϕ_1	0.626	ϕ_2	0.0	ϕ_3	0.184
				(0.083)				(0.086)
Gota	(2,0)	1.0	ϕ_1	0.591	ϕ_2	-0.274		
				(0.079)		(0.086)		
Espanola	(1,1)	0.0	ϕ_1	0.963	θ_1	0.537		
				(0.016)		(0.051)		
Temp	(0,2)	1.0	θ_1	-0.115	θ_2	-0.202		
				(0.063)		(0.057)		
Precip	(0,0)	0.0						
Sunyr	(9,0)	1.0	ϕ_1	1.219	ϕ_2	-0.508	ϕ_9	0.232
				(0.060)		(0.056)		(0.029)
Minimum	(2,1)	-0.778	ϕ_1	1.254	ϕ_2	-0.279	θ_1	0.842
		(0.316)		(0.060)		(0.051)		(0.049)
Snake	(3,0)	1.0	ϕ_1	0.352	ϕ_2	0.093	ϕ_3	0.100
				(0.039)		(0.041)		(0.039)
Exshaw	(1,1)	1.0	ϕ_1	0.725	θ_1	0.395		
				(0.067)		(0.090)		
Naramata	(2,0)	1.0	ϕ_1	0.196	ϕ_2	0.131		
				(0.044)		(0.044)		
Dell	(2,0)	1.0	ϕ_1	0.367	ϕ_2	0.185		
				(0.039)		(0.039)		
Lakeview	(3,0)	0.717	ϕ_1	0.525	ϕ_2	0.0	ϕ_3	0.143
		(0.130)		(0.038)				(0.039)
Ninemile	(2,1)	0.684	ϕ_1	1.225	ϕ_2	-0.274	θ_1	0.850
		(0.060)		(0.063)		(0.047)		(0.049)
Eaglecol	(2,1)	0.624	ϕ_1	1.156	ϕ_2	-0.237	θ_1	0.693
		(0.054)		(0.114)		(0.082)		(0.103)
Navajo	(1,1)	1.0	ϕ_1	0.683	θ_1	0.424		
				(0.082)		(0.103)		
Bryce	(1,0)	1.366	ϕ_1	0.598				
		(0.107)		(0.033)				
Tioga	(1,0)	1.458	ϕ_1	0.556				
		(0.098)		(0.033)				
Bigcone	(2,0)	1.0	ϕ_1	0.375	ϕ_2	0.159		
				(0.044)		(0.044)		
Whitemtn	(1,1)	1.414	ϕ_1	0.641	θ_1	0.408		
		(0.061)		(0.086)		(0.104)		

*The parenthetical values are standard errors (SE's).

distribution function (CDF) of $\bar{R}*_N$ for a white noise process and in general any ARMA model is analytically intractable. However, by simulation it is possible to determine as accurately as is desired for practical purposes the CDF for $\bar{R}*_N$. Because both $\bar{R}*_N$ and K are functions of N, their CDF's are defined for a particular length of series N. The CDF for $\bar{R}*_N$ is

$$F = F(r;N,\phi,\theta) = Pr(\bar{R}*_N \leq r) \qquad [10.6.1]$$

where N is the length of each individual time series, ϕ is the set of known AR parameters, θ is the set of known MA parameters, and r is any possible value of $\bar{R}*_N$.

When simulating a time series of length N, it is recommended to employ the improved simulation techniques of Chapter 9. In this section, WASIM1 from Section 9.3 is utilized for the ARMA $(0,q)$ models, while WASIM2 from Section 9.4 is used with the ARMA$(p,0)$ and ARMA(p,q) processes. Because the RAR or K is independent of the variance of the innovations, any value of σ_a^2 may be used. Consequently, it is simplest to set $\sigma_a^2 = 1$ and hence to assume that the residuals are NID(0,1).

Suppose that $\bar{N}$ simulations of length N are generated for a specific ARMA model and the $\bar{N}$ RAR's given by $\bar{R}*_{N1}, \bar{R}*_{N2}, \ldots, \bar{R}*_{N\bar{N}}$, are calculated for the $\bar{N}$ simulated series, respectively. If the sample of RAR is reordered such that $\bar{R}*_{N(1)} \leq \bar{R}*_{N(2)} \leq \cdots \leq \bar{R}*_{N(\bar{N})}$, it is known that the MLE of F is given by the ECDF (Gnedenko, 1968, pp. 444-451):

$$\begin{aligned} F_{\bar{N}} &= F_{\bar{N}}(r;N,\phi,\theta) = 0\,, \quad r \leq \bar{R}*_{N(1)} \\ F_{\bar{N}} &= F_{\bar{N}}(r;N,\phi,\theta) = k/N\,, \quad \bar{R}*_{N(k)} < r \leq \bar{R}*_{N(k+1)} \qquad [10.6.2] \\ F_{\bar{N}} &= F_{\bar{N}}(r;N,\phi,\theta) = 1\,, \quad r > \bar{R}*_{N(\bar{N})} \end{aligned}$$

The Kolmogorov theorem (Gnedenko, 1968, p. 450) can be used to obtain confidence intervals for $F_{\bar{N}}$ and to indicate the number of samples $\bar{N}$ necessary to guarantee a prescribed accuracy. This theorem states that if $\bar{N}$ is moderately large (it has been shown that $\bar{N} > 100$ is adequate), then

$$Pr(\max_r |F_{\bar{N}} - F| < \varepsilon/N^{1/2}) \simeq K(\varepsilon) \qquad [10.6.3]$$

where

$$K(\varepsilon) = 0\,, \quad \varepsilon \leq 0$$

$$K(\varepsilon) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\varepsilon^2}\,, \quad \varepsilon > 0$$

For example, when $\varepsilon = 1.63$, then $K(\varepsilon) = 0.99$. If $\bar{N} = 10^4$ simulations are done for a series of length N, then by Kolmogorov's theorem, all the values of $F_{\bar{N}}$ are accurate to at least within 0.0163 with a probability of 0.99.

In actual simulation studies, it is useful to examine the convergence of $F_{\bar{N}}$ by printing out a summary of the ECDF for increasing values of $\bar{N}$ (such as $\bar{N} = 100{,}200{,}500{,}1000{,}2000, \cdots$) until sufficient accuracy has been obtained. To curtail the computer time required in simulations, there are efficient algorithms available called "quicksorts" (Knuth, 1973) for ordering the sample values for the RAR.

If simulation studies are done for $\bar{R}^*_N$, the ECDF for K can be obtained from the transformation

$$K = \log\bar{R}^*_N/\log(N/2) \qquad [10.6.4]$$

Alternatively, when the ECDF for K is known, the ECDF for $\bar{R}^*_N$ can be calculated by substituting each value of K into

$$\bar{R}^*_N = (N/2)^K \qquad [10.6.5]$$

Representative ECDF's are given in Appendix A10.1 for the simulations carried out in this section. In Table A10.1.1, the ECDF of K is shown for various values of N for white noise that is NID$(0,\sigma_a^2)$. For each value of N (i.e., each row) in that table, an ECDF is determined using $\bar{N} = 10^4$ samples of length N. By substituting all values for K in this table into [10.6.5] the ECDF for the RAR can be found for each value of N.

When a particular time series is modelled by an ARMA model other than white noise, the ECDF for either $\bar{R}^*_N$ or K can be calculated by simulation for each desired value of N. Table A10.1.2 lists the ECDF of K for different values of N for a Markov process in [3.2.1] with $\phi_1 = 0.4$. By utilizing the transformation in [10.6.5] for each entry in this table, the ECDF's for $\bar{R}^*_N$ for the Markov model can be found and are shown in Table A10.1.3. Because of the transformation in [10.6.5], it is sufficient to simply have a table for either K or $\bar{R}^*_N$.

The tables of various ECDF's for different types of ARMA models are listed in the appendix of the microfiche version of the paper by Hipel and McLeod (1978a). In particular, results are given for white noise as well as Markov models with $\phi_1 = 0.1, 0.2, \ldots, 0.9$. In all of the tables, for a particular value of N the number of samples $\bar{N}$ simulated is 10^4.

For a particular ARMA model, the ECDF can be used to make inferences about $\bar{R}^*_N$ or equivalently K. For instance, the 95% confidence interval for $\bar{R}^*_N$ with $N = 100$ for a Markov or AR(1) model with $\phi_1 = 0.4$ can be determined by utilizing Table A10.1.3. Opposite $N = 100$, select the values of $\bar{R}^*_N$ below the 0.025 and 0.975 quantiles. The 95% confidence interval for the RAR is then 9.85 - 24.02. By substituting these interval limits into [10.6.4], the 95% confidence interval for K is 0.585 - 0.813. This confidence interval for K is also confirmed by referring to the appropriate entries of Table A10.1.2 opposite $N = 100$.

The ECDF tables illustrate certain properties of the RAR or K. For example, an examination of the median for K for white noise below the 0.500 quantile in Table A10.1.1, definitely shows that K slowly decreases asymptotically toward 0.5 with increasing N and is consequently a function of N. Because of this, a separate ECDF must be developed for each value of N for a specified process. Note that the median values for K in Table A10.1.1 are almost identical with the values of K tabulated in Table 10.5.2. These latter values of K are calculated by using [10.6.4] when the exact theoretical expected values of $\bar{R}^*_N$ are found from a formula given by Anis and Lloyd (1976) and also by employing simulation techniques to estimate $E(K)$. As can be seen from a perusal of Table 10.5.2, the expected value of K is obviously a function of N and decreases in magnitude with increasing N.

It can be proven theoretically that for any ARMA process the RAR or K is a function of the time series length N and the AR and MA parameters (Hipel, 1975, Appendix B). This fact is confirmed by the ECDF for the RAR for the Markov process with $\phi_1 = 0.4$ in Tables A10.1.2 and A10.1.3. It can be seen that the median and all other values of the RAR at any quantile for all of the Markov models increase in value for increasing N. When one compares the results with other Markov models given by Hipel and McLeod (1978a, microfiche appendix), the distribution of $\bar{R}^*_N$ or K is also a function of the value of the AR parameter ϕ_1.

10.6.4 Preservation of the RAR and K by ARMA Models

By employing the ECDF of the RAR or K in conjunction with a specified statistical test it is now shown that ARMA models do preserve the historically observed Hurst statistics. Because the Hurst coefficient K is widely cited in the literature, the research results for this statistic are described. However, K and $\bar{R}^*_N$ are connected by the simple transformation given in [10.6.5], and, therefore, preservation of either statistic automatically implies retention of the other by an ARMA model.

The ARMA models fitted to 23 geophysical time series ranging in length from $N = 96$ to $N = 1164$ are listed in Table 10.6.3. For exactly the same time series length N as the historical data, 10^4 simulations are done for each model to determine the ECDF of K, or equivalently $\bar{R}^*_N$. The probability p_i of having K for the ith model greater than the K calculated for the ith historical series is determined from the ith ECDF as

$$Pr(K > K_i^{obs} | \text{model}) = p_i \qquad [10.6.6]$$

where K_i^{obs} is the K value calculated for the ith observed historical time series. If the chosen ARMA model is correct, then, by definition, p_i would be uniformly distributed on (0,1). For k time series it can be shown (Fisher, 1970, p. 99) that

$$-2\sum_{i=1}^{k} \ln p_i - \chi^2_{2k} \qquad [10.6.7]$$

Significance testing can be done by using [10.6.7] to determine whether the observed Hurst coefficient or the RAR is preserved by ARMA models. The test could fail if the incorrect model were fitted to the data (for example, if the Ogden data were incorrectly modelled by an AR(1) process with $\phi_1 = 0.4$) or if ARMA models do not retain the Hurst K. Careful model selection was done, thereby largely eliminating the former reason for test failure. If it is thought (as was suggested by Mandelbrot and Wallis (1968)) that the observed K is larger than that implied by an appropriate Brownian domain model, then a one tailed rather than a two tailed test may be performed.

The results of the χ^2 test in [10.6.7] for the 23 geophysical phenomena confirm that there is no evidence that the observed K's, or equivalently the RAR's, are not adequately preserved by the fitted ARMA models. Table 10.6.4 summarizes the information used in the test. The observed Hurst coefficient, $E(K)$ from the simulations and the p_i value are listed for each of the time series. In Table 10.6.5, it can be seen that the calculated χ^2 value from [10.6.7] is not significant at the 5% level of significance for the 23 time series for either a one sided or a two sided test. Therefore, on the basis of the given information, ARMA models do statistically preserve K or the RAR when considering all the time series. Furthermore, when the set of annual riverflows, miscellaneous data, and tree ring indicies are inspected individually, it can be seen from Table 10.6.5 that ARMA models preserve the historical Hurst statistics for all three cases.

Table 10.6.4. Geophysical time series calculations.

Code Names	N's	Observed K's	ARMA Model $E(K)$'s	p_i's
Mstouis	96	0.648	0.667	0.624
Neumunas	132	0.660	0.649	0.420
Danube	120	0.633	0.613	0.534
Rhine	150	0.614	0.609	0.468
Ogden	97	0.894	0.832	0.149
Gota	150	0.689	0.659	0.283
Espanola	350	0.855	0.877	0.674
Temperature	255	0.694	0.646	0.157
Precip	100	0.618	0.610	0.434
Sunspot numbers	163	0.723	0.768	0.728
Minimum	848	0.815	0.786	0.264
Snake	669	0.687	0.693	0.559
Exshaw	506	0.637	0.702	0.938
Naramata	515	0.595	0.649	0.905
Dell	655	0.687	0.694	0.569
Lakeview	544	0.706	0.729	0.709
Ninemile	771	0.740	0.726	0.378
Eaglecol	858	0.645	0.747	0.995
Navajo	700	0.653	0.670	0.660
Bruce	625	0.732	0.698	0.203
Tioga	661	0.701	0.687	0.362
Bigcone	509	0.611	0.695	0.981
Whitemtn	1164	0.695	0.648	0.095

Table 10.6.5. Results of the χ^2 test for the geophysical time series.

Data Sets	Degrees of	$-2\ln\sum p_i$
	Freedom	
Riverflows	12	11.78
Miscellaneous	10	9.46
Tree rings	24	16.08
Total	46	37.32

In Table 10.6.4, the average of the observed K's is calculated to be 0.693 with a standard deviation of 0.076. The $E(K)$ from the simulations has an average of 0.698 with a standard deviation of 0.068. The average of the observed K is, therefore, slightly less than that for the simulated case, but this difference is not statistically different.

If the results of the RAR had been given rather than K, only columns 3 and 4 of Table 10.6.4 would be different, due to the transformation in [10.6.7]. The p_i values and the results of the χ^2 test in Table 10.6.5 would be identical. Therefore, preservation of either K or $\bar{R}^*_N$ infers retention of the other statistic by ARMA models.

10.7 ESTIMATES OF THE HURST COEFFICIENT

Different estimators are available for estimating the Hurst coefficient. The purpose of this section is to compare these estimates for the 23 annual geophysical time series given in Tables 10.6.1 and 10.6.2.

From empirical studies of approximately 690 geophysical time series, Hurst (1951, 1956) found the RAR to vary as

$$\bar{R}^*_N \propto N^h \qquad [10.7.1]$$

where h is a constant often referred to as the generalized Hurst coefficient. The above equation can be written in the general form

$$\bar{R}^*_N = aN^h \qquad [10.7.2]$$

where a is a coefficient. Hurst assumed the coefficient a to have a value of $(1/2)^h$ and then estimated h by K in [10.6.4].

Siddiqui (1976) has employed the functional central limit theorem and the theory of Brownian motion to derive many statistical formulae that may be of interest to hydrologists. Of particular importance is the asymptotic result for calculating $E(\bar{R}^*)$ for ARMA processes. This formula is given as

$$E(\bar{R}^*) \simeq a'N^{1/2} \qquad [10.7.3]$$

where

$$a' = 1.2533\gamma_0^{-1/2}\left(1 - \sum_{i=1}^{q}\theta_i\right) \Big/ \left(1 - \sum_{i=1}^{p}\phi_i\right)$$

and γ_0 is the theoretical autocovariance function at lag 0 that is evaluated by using the algorithm in Appendix A3.2 with $\sigma_a^2 = 1$, θ_i is the ith MA parameter and ϕ_i is the ith AR parameter. If the random variables are IID, a special case of [10.7.3] that was previously derived by Feller (1951) is

$$E(\bar{R}^*_N) \simeq 1.2533N^{1/2} \qquad [10.7.4]$$

By comparing [10.7.3] and [10.7.2], a possible alternative method of evaluating h may be to employ the equation

$$\bar{R}*_N = a'N^{SH} \quad [10.7.5]$$

where SH is Siddiqui's estimate of the generalized Hurst coefficient h. When logarithms are taken of [10.7.5], Siddiqui's estimate for h is (Siddiqui, 1976)

$$SH = (\log\bar{R}*_N - \log a')(\log N)^{-1} \quad [10.7.6]$$

It should be noted that due to the way Hurst (1951, 1956) and Siddiqui (1976) calculate the coefficient a in [10.7.2], the Hurst coefficient K and the Siddiqui coefficient SH are in fact two different statistics. Nevertheless, as was suggested by Siddiqui (1976), it may be of interest to determine whether h exhibits the Hurst phenomenon if the estimate SH is employed. Accordingly, for the 23 geophysical time series given in Tables 10.6.1 and 10.6.2 the K and SH statistics are compared.

Table 10.6.3 lists the ARMA models fitted to the 23 time series. If a Box-Cox transformation is included in a model, then K and SH are calculated for the transformed series to which the model is fit. This is because the formula for calculating SH in [10.7.6] does not have the capability of incorporating a Box-Cox transformation in order to get an estimate of SH for the untransformed data. Table 10.7.1 displays the values of K and SH that are calculated for each time series by using [10.6.4] and [10.7.6], respectively. Notice that the entries for K in Table 10.7.1 differ from the K values in Table 10.6.4 wherever the data used in Table 10.7.1 have been transformed by a Box-Cox transformation.

An examination of Table 10.7.1 reveals that in all cases except three, the value of SH is less than K for the corresponding time series. The K statistic has an arithmetic mean of 0.701 with a standard deviation of 0.084. However, the mean of the SH statistic is 0.660 and possesses a standard deviation of 0.131. The mean value of SH is, therefore, well within 2 standard deviations of 0.500.

Another technique to estimate h can be found by comparing [10.7.4] and [10.7.2]. Accordingly, Gomide (1975) suggests the following equation to evaluate h:

$$\bar{R}*_N = 1.2533N^{YH} \quad [10.7.7]$$

where YH is Gomide's estimate of the generalized Hurst coefficient h. By taking logarithms of [10.7.7], Gomide's estimate of h is

$$YH = (\log\bar{R}*_N - \log 1.2533)(\log N)^{-1} \quad [10.7.8]$$

When [10.7.8] is utilized to estimate the Hurst coefficient, Gomide (1975) obtains an average value for YH of 0.57 for the 690 series considered by Hurst (1951, 1956). On the other hand, Hurst (1951, 1956) calculated K to have an average of 0.73 for the 690 series. Therefore, lower values are obtained for the Hurst coefficient h if YH is employed rather than K.

Table 10.7.1 lists the values of YH for the same 23 geophysical time series that are considered for SH. Therefore, if a Box-Cox transformation is included with an ARMA model in Table 10.6.3, then YH is determined for the transformed series to which the model is fit. Obviously, because YH, as calculated in [10.7.8], is not a function of the ARMA model parameters, it is not, in general, necessary to consider the transformed series. However, the aforementioned procedure is adopted so that appropriate comparisons can be formulated for the three estimates given in Table 10.7.1.

Table 10.7.1. Estimates of the Hurst coefficient.

Code Names	*K*'s	*SH*'s	*YH*'s
Mstouis	0.648	0.451	0.500
Neumunas	0.677	0.499	0.535
Danube	0.633	0.495	0.495
Rhine	0.614	0.484	0.484
Ogden	0.894	0.436	0.709
Gota	0.689	0.504	0.549
Espanola	0.928	0.455	0.779
Temp	0.694	0.521	0.567
Precip	0.615	0.473	0.473
Sunyr	0.723	0.570	0.580
Minimum	0.817	0.462	0.699
Snake	0.687	0.475	0.579
Exshaw	0.637	0.420	0.530
Naramata	0.595	0.435	0.492
Dell	0.687	0.475	0.579
Lakeview	0.703	0.499	0.590
Ninemile	0.727	0.466	0.617
Eaglecol	0.761	0.485	0.650
Navajo	0.653	0.468	0.550
Bryce	0.734	0.513	0.620
Tioga	0.704	0.498	0.594
Bigcone	0.611	0.404	0.507
Whitment	0.695	0.530	0.595

A perusal of Table 10.7.1 shows that for each time series the values of both *SH* and *YH* is consistently less than the magnitude of *K*. For the series to which white noise models are fit in Table 10.6.3 (i.e., Danube, Rhine and Precip), the values of *YH* and *SH* in Table 10.7.1 are equivalent. However, for all the other data sets the magnitudes of *SH* are less than *YH*. The mean of the 23 *YH* values is 0.577 with a standard deviation of 0.078. The *YH* statistic is within one standard deviation of 0.500. Therefore, it can perhaps be argued that for the data considered, the Hurst phenomenon is not significant for the *YH* statistic. A similar argument can be made for the *SH* estimate of *h*.

10.8 CONCLUSIONS

The pursuit of possible explanations to solve the riddle of the Hurst phenomenon has stimulated decades of valuable research by both hydrologists and statisticians. The Hurst researchers are analogous to the inquisitive archaelogists of the 19th and early 20th centuries who sought to find the treasures of the ancient Egyptians in long forgotten temples, pyramids and tombs. Like the archaelogists, during their search the Hurst scientists have unearthed many valuable treasures that have attracted the world-wide attention of their colleagues. However, the main treasure find is the one described in Section 10.6. In that section, ARMA models are shown to preserve statistically the observed RAR and *K* when fitted to a variety of geophysical time series. In other words, the fitted ARMA models indirectly account for the measured Hurst statistics, which are usually significantly larger than 0.5 (see Table 10.6.4). Because important stochastic characteristics of hydrologic time series are retained by ARMA models, this should give engineers confidence in water resource projects that are designed with the aid of simulation techniques. In particular, the RAR statistic is directly related to storage problems, and this makes ARMA models desirable for reservoir design, operation, and evaluation.

Besides the main solution to the Hurst riddle given in Section 10.6, many other interesting discoveries have been made. In addition to the Hurst coefficient K defined in [10.3.4], other coefficients have been suggested to model the generalized Hurst coefficient h given in [10.3.2]. For example, Gomide (1975), Siddiqui (1976), Anis and Lloyd (1976) and Wallis and Matalas (1970) proposed alternative procedures to model h. One of the major reasons for developing alternative exponents to K was to produce a coefficient that would reach its limiting value of 0.5 more quickly than K would. Nevertheless, it must be borne in mind that the definition of the Hurst phenomenon is based on a comparison of the value of K in small and moderate sample sizes to its large sample value of 0.5. If the empirical, or theoretical, value of another estimate of h is compared for finite time series length to its asymptotic magnitude of 0.5, the Hurst phenomenon should probably be redefined in terms of that statistic. However, because of the inherent statistical properties of the RAR, it is recommended that future research primarily be devoted to the study of this statistic and that less emphasis be put on the various definitions of the Hurst coefficient. Some interesting insights into problems related to the Hurst phenomenon are provided by Klemes and Klemes (1988). Further research into the Hurst phenomenon and long-range dependence is provided by Bhattacharya et al. (1983) and Poveda and Mesa (1988) while Beran (1992) carries out a partial survey of long-range dependence research. Kunsch (1986) provides an approach for discriminating between monotonic trends and long-range dependence. Finally, Cox (1991) links non-linearity and time irreversibility with long-range dependence.

Feller (1951) proved that the asymptotic formula for the expected value of the adjusted range in [10.3.5] is valid for IID random variables. As is shown in [10.3.6] for large samples, Feller's equation is also correct for the expected value of the RAR for IID summands. The exact analytical expression for the expected value of the RAR for NID summands was derived by Anis and Lloyd (1976) and is written in [10.3.15]. For finite samples, the simulation and analytical results of Table 10.5.1 indicate that the expected values of the RAR and hence K are functions of the sample size but are virtually independent of the underlying distribution for IID summands. Accordingly, it has been suggested that the Hurst phenomenon could be explained by a combination of transcience and autocorrelation (Wallis and O'Connell, 1973). This implies that perhaps either a short memory or a long memory model that takes into account the autocorrelation structure of a time series may explain the Hurst phenomenon. Perhaps a better way to phrase this is that if a given stochastic model, that is fit to a given data set, preserves the important historical statistics such as the RAR and K, then that model may indirectly account for the Hurst phenomenon. Therefore, it can be argued that a resolution to the controversies related to the Hurst phenomenon boils down to determining stochastic models that preserve the RAR, as well as other relevant historical statistics.

If a stochastic model is to retain the historical statistical characteristics of a time series, then the model must provide a good statistical fit to the data. This can be accomplished in practice by following the identification, estimation, and diagnostic check stages of model construction described in Part III of the book. For long memory FGN processes the authors have developed an efficient estimation procedure using the method of maximum likelihood (Section 10.4.3), and a technique for calculating the model residuals so that they can be tested by appropriate diagnostic checks (Section 10.4.4). Moreover, in Section 10.4.5 a method is given for calculating one step ahead MMSE forecasts for a FGN model. Finally, in Section 10.4.6 a technique is presented for exactly simulating FGN such that the synthetic traces will lie outside the Brownian domain for the parameter H in the range $0.5 < H < 1$.

Short memory models provide an alternative approach to FGN processes for modelling hydrological time series. In particular, the ARMA family of short memory models possesses great potential for widespread applications to water resource as well as other geophysical and environmental problems. Klemes et al. (1981) maintain that given the socio-economic and hydrologic data usually available for reservoir planning and design, the replacement of short memory models with long memory ones in reservoir analyses, cannot be objectively justified.

A statistical approach for discriminating between short and long memory models is to use the AIC of Section 6.3. The AIC provides a means of model discrimination based on the principles of good statistical fit and parsimony of the model parameters. For the six annual riverflow time series considered in Section 10.4.7 the results of Table 10.4.4 show that in all six cases the AIC chooses the best fitting ARMA model in preference to the FGN process. Although there may be certain situations where the FGN model is appropriate to use, the inherent inflexibility of a FGN process may limit the use of this model in many types of practical applications. Rather than allowing for a choice of the required number of model parameters to use in a given situation as is done in ARMA modelling, the FGN model is always restricted to just three parameters (i.e., the mean, the variance, and H).

By adhering to the model construction stages of Part III, it is a straightforward procedure to develop an appropriate ARMA to describe a particular time series. If the phenomenon being modelled has been influenced significantly by external interventions, these effects can be incorporated into the model using the intervention model of Chapter 19. By employing Monte Carlo techniques, the ECDF's of statistics such as the RAR or K can be developed to any desired accuracy, as shown in Section 10.6.3. The ECDF's are used in conjunction with a specified statistical test to check for the preservation of historical statistics in Section 10.6.4. This testing procedure can be used to check for the retention of any observed statistics by ARMA or by other types of stochastic models. Tsay (1992), for example, employs a similar approach for investigating the reproducibility of historical statistics by fitted models.

Besides considering Hurst's estimate K of the coefficient h, it is possible to entertain other types of estimates as explained in Section 10.7. For the 23 natural time series listed in Tables 10.6.1 and 10.6.2, the Siddiqui coefficient SH (Siddiqui, 1976) and Gomide's statistic YH (Gomide, 1975) possess a mean value less than K. By examining the standard deviations of the YH and SH statistics the Hurst phenomenon is seen to be less pronounced for these estimates than it is for K.

If one wishes to consider fitting a long memory model to a specified time series, one may wish to entertain the fractional ARMA or FARMA model as an alternative to FGN. This model is more flexible to use than the long memory FGN model because the number of model parameters is not fixed. In fact, as shown in the next chapter, the FARMA model is a direct extension of ARMA and ARIMA models.

APPENDIX A10.1

REPRESENTATIVE

EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTIONS

(ECDF's)

FOR HURST STATISTICS

The three tables presented in this appendix contain ECDF's for simulation studies explained in Section 10.6.3. More specifically, for a range of lengths N of simulated sequences the following three sets of ECDF's are given:

Table A10.1.1. ECDF's of K for a NID$(0,\sigma_a^2)$ process.

Table A10.1.2. ECDF's of K for a Markov process with $\phi_1 = 0.4$.

Table A10.1.3. ECDF's of $\bar{R}^*_N$ for a Markov process $\phi_1 = 0.4$.

Table A10.1.1. ECDF's of *K* for a NID(0,σ_a^2) process.*

Value of *N*	0.005	0.010	0.025	0.050	0.100	0.200	0.300	0.400
5	0.260	0.294	0.354	0.41?	0.485	0.474	0.634	0.677
10	0.355	0.380	0.422	0.450	0.506	0.566	0.611	0.648
15	0.390	0.414	0.445	0.478	0.518	0.570	0.605	0.637
20	0.406	0.426	0.459	0.486	0.520	0.565	0.600	0.628
25	0.414	0.436	0.466	0.497	0.527	0.569	0.509	0.626
30	0.429	0.445	0.473	0.501	0.531	0.570	0.598	0.622
35	0.436	0.453	0.482	0.505	0.534	0.569	0.596	0.619
40	0.446	0.463	0.488	0.511	0.537	0.570	0.596	0.617
45	0.447	0.464	0.487	0.519	0.537	0.570	0.594	0.615
50	0.453	0.468	0.489	0.519	0.537	0.570	0.593	0.613
60	0.461	0.473	0.494	0.515	0.538	0.569	0.591	0.610
70	0.460	0.474	0.498	0.515	0.538	0.568	0.589	0.607
80	0.461	0.475	0.499	0.518	0.450	0.568	0.588	0.606
90	0.464	0.483	0.501	0.519	0.540	0.569	0.588	0.604
100	0.466	0.480	0.499	0.520	0.541	0.566	0.585	0.601
125	0.474	0.489	0.507	0.523	0.542	0.565	0.583	0.598
150	0.477	0.489	0.506	0.523	0.542	0.564	0.581	0.596
175	0.476	0.488	0.508	0.523	0.540	0.563	0.579	0.594
200	0.484	0.494	0.509	0.525	0.542	0.564	0.479	0.593
500	0.491	0.497	0.512	0.524	0.538	0.556	0.559	0.580
1000	0.494	0.502	0.514	0.525	0.537	0.552	0.594	0.573

*Table continues on opposite page.

(Table A10.1.1 continued.)

. . .

Quantile 0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995
0.714	0.748	0.801	0.854	0.904	0.932	0.948	0.961	0.967
0.685	0.720	0.755	0.791	0.836	0.872	0.895	0.917	0.932
0.668	0.696	0.724	0.757	0.800	0.832	0.855	0.882	0.897
0.655	0.681	0.708	0.738	0.775	0.804	0.827	0.852	0.870
0.651	0.674	0.698	0.726	0.762	0.790	0.813	0.836	0.849
0.644	0.666	0.689	0.716	0.750	0.777	0.799	0.822	0.835
0.640	0.661	0.683	0.709	0.743	0.768	0.790	0.809	0.823
0.637	0.656	0.679	0.702	0.734	0.759	0.780	0.801	0.815
0.634	0.654	0.675	0.699	0.730	0.753	0.774	0.791	0.806
0.632	0.650	0.669	0.692	0.722	0.748	0.766	0.786	0.801
0.627	0.645	0.664	0.684	0.714	0.735	0.751	0.772	0.786
0.623	0.641	0.658	0.678	0.706	0.728	0.746	0.765	0.777
0.622	0.638	0.655	0.675	0.702	0.723	0.740	0.757	0.769
0.619	0.635	0.652	0.671	0.696	0.716	0.723	0.754	0.758
0.616	0.632	0.648	0.666	0.691	0.711	0.728	0.746	0.761
0.613	0.627	0.642	0.660	0.684	0.703	0.718	0.732	0.744
0.609	0.624	0.638	0.555	0.678	0.697	0.713	0.730	0.740
0.607	0.620	0.634	0.650	0.672	0.689	0.704	0.720	0.729
0.606	0.619	0.632	0.647	0.668	0.685	0.698	0.713	0.721
0.591	0.602	0.613	0.626	0.643	0.657	0.669	0.683	0.693
0.583	0.592	0.602	0.614	0.630	0.642	0.653	0.664	0.671

. . .

Table A10.1.2. ECDF's of K for a Markov process with $\phi_1 = 0.4$.*

Value of N	0.005	0.010	0.025	0.050	0.100	0.200	0.300	0.400
5	0.236	0.325	0.407	0.476	0.551	0.635	0.636	0.724
10	0.401	0.442	0.494	0.537	0.590	0.660	0.703	0.746
15	0.470	0.494	0.536	0.503	0.612	0.666	0.705	0.738
20	0.435	0.507	0.545	0.580	0.618	0.665	0.700	0.730
25	0.495	0.512	0.555	0.586	0.622	0.667	0.700	0.726
30	0.514	0.530	0.566	0.594	0.628	0.667	0.697	0.721
35	0.523	0.544	0.573	0.593	0.627	0.664	0.693	0.716
40	0.530	0.550	0.577	0.603	0.631	0.666	0.591	0.714
45	0.536	0.551	0.578	0.603	0.631	0.664	0.690	0.711
50	0.540	0.555	0.532	0.692	0.629	0.662	0.688	0.707
60	0.542	0.559	0.582	0.602	0.628	0.661	0.683	0.703
70	0.546	0.553	0.583	0.603	0.626	0.658	0.679	0.698
80	0.542	0.559	0.583	0.605	0.628	0.655	0.677	0.695
90	0.549	0.562	0.585	0.604	0.626	0.656	0.676	0.692
100	0.550	0.563	0.585	0.605	0.625	0.651	0.671	0.688
125	0.558	0.568	0.588	0.605	0.624	0.649	0.667	0.683
150	0.555	0.564	0.534	0.602	0.622	0.646	0.663	0.673
175	0.552	0.567	0.586	0.601	0.619	0.642	0.660	0.674
200	0.559	0.570	0.586	0.602	0.620	0.642	0.658	0.672

*Table continues on opposite page.

(Table A10.1.2 continued.)

. . .

Quantile								
0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995
0.763	0.810	0.851	0.835	0.922	0.943	0.958	0.967	0.971
0.777	0.806	0.833	0.863	0.894	0.916	0.932	0.948	0.958
0.766	0.792	0.819	0.845	0.877	0.901	0.918	0.931	0.941
0.756	0.778	0.804	0.830	0.862	0.884	0.900	0.916	0.927
0.750	0.773	0.795	0.820	0.851	0.872	0.889	0.906	0.914
0.744	0.765	0.787	0.811	0.841	0.863	0.880	0.898	0.905
0.737	0.759	0.780	0.803	0.833	0.854	0.872	0.887	0.898
0.734	0.754	0.774	0.796	0.826	0.847	0.864	0.830	0.893
0.731	0.749	0.770	0.793	0.821	0.841	0.857	0.873	0.885
0.726	0.744	0.763	0.785	0.813	0.836	0.852	0.870	0.879
0.720	0.738	0.756	0.776	0.803	0.823	0.838	0.855	0.866
0.715	0.731	0.749	0.769	0.795	0.815	0.831	0.849	0.859
0.712	0.728	0.745	0.765	0.790	0.809	0.826	0.841	0.851
0.703	0.724	0.740	0.759	0.784	0.803	0.818	0.836	0.847
0.704	0.719	0.735	0.753	0.778	0.797	0.813	0.829	0.840
0.697	0.712	0.727	0.745	0.763	0.786	0.801	0.816	0.825
0.692	0.706	0.721	0.738	0.761	0.780	0.795	0.810	0.813
0.688	0.701	0.715	0.732	0.752	0.771	0.785	0.799	0.809
0.685	0.698	0.712	0.727	0.748	0.764	0.777	0.792	0.801

. . .

Table A10.1.3. ECDF's of $\bar{R}^*_n$ for a Markov process with $\phi_1 = 0.4$.*

Value of N	0.005	0.010	0.025	0.050	0.100	0.200	0.300	0.400
5	1.30	1.35	1.45	1.55	1.66	1.79	1.88	1.94
10	1.91	2.04	2.21	2.37	2.59	2.89	3.13	3.32
15	2.58	2.71	2.95	3.14	3.43	3.82	4.14	4.42
20	3.05	3.22	3.51	3.80	4.15	4.62	5.01	5.37
25	3.49	3.73	4.06	4.39	4.82	5.39	5.86	6.26
30	4.02	4.20	4.63	4.99	5.48	6.03	6.60	7.05
35	4.47	4.75	5.15	5.54	6.01	6.69	7.27	7.76
40	4.89	5.19	5.63	6.09	6.63	7.35	7.93	8.50
45	5.30	5.55	6.06	6.53	7.13	7.91	8.57	9.15
50	5.68	5.97	6.51	6.94	7.58	8.41	9.15	9.75
60	6.33	6.70	7.23	7.75	8.47	9.46	10.21	10.92
70	6.96	7.26	7.95	8.54	9.27	10.38	11.17	11.94
80	7.38	7.85	8.60	9.30	10.13	11.20	12.16	12.98
90	8.08	8.49	9.26	9.98	10.86	12.13	13.00	13.91
100	8.59	9.06	9.85	10.68	11.52	12.78	13.83	14.74
125	10.06	10.48	11.38	12.21	13.23	14.65	15.74	16.85
150	10.99	11.43	12.47	13.47	14.65	16.28	17.52	18.71
175	11.81	12.60	13.72	14.66	15.91	17.65	19.09	20.39
200	13.15	13.77	14.84	15.99	17.39	19.21	20.71	22.10

*Table continues on opposite page.

(Table A10.1.3 continued.)

Quantile								
0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995
2.01	2.10	2.18	2.25	2.33	2.37	2.40	2.43	2.43
3.49	3.66	3.82	4.01	4.22	4.37	4.48	4.60	4.67
4.68	4.94	5.20	5.49	5.85	6.14	6.36	6.53	6.65
5.70	6.00	6.37	6.77	7.28	7.65	7.95	8.24	8.45
6.64	7.04	7.46	7.92	8.57	9.05	9.45	9.35	10.05
7.49	7.95	8.44	8.98	9.75	10.34	10.85	11.38	11.64
8.25	8.79	9.32	9.96	10.84	11.54	12.14	12.68	13.08
9.02	9.56	10.15	10.86	11.87	12.65	13.29	13.98	14.53
9.72	10.31	19.99	11.80	12.87	13.73	14.43	15.17	15.72
10.34	10.95	11.65	12.52	13.71	14.73	15.54	16.45	16.95
11.57	12.30	13.07	14.03	15.36	16.43	17.31	18.32	19.04
12.70	13.47	14.32	15.41	16.90	18.16	19.22	20.44	21.23
13.82	14.65	15.62	16.80	18.44	19.81	21.02	22.28	23.05
14.80	15.71	16.72	17.97	19.76	21.27	22.54	24.09	25.18
15.68	16.65	17.76	19.01	20.97	22.64	24.02	25.64	26.72
17.88	18.97	20.21	21.79	23.97	25.77	27.42	29.26	30.27
19.86	21.12	22.53	24.24	26.72	28.98	30.92	33.05	34.25
21.71	23.00	24.50	26.42	28.91	31.44	33.51	35.62	37.22
23.42	24.89	26.54	28.46	31.33	33.69	35.81	38.40	39.95

PROBLEMS

10.1 Read Hurst's (1951, 1956) original papers about his work in long term storage. Summarize what he did and comment upon his abilities as an engineer and a statistician.

10.2 In Sections 10.2 and 10.3.1, statistics are defined for studying long term storage problems. Suggest some statistics for examining short term storage problems in reservoir design.

10.3 What do you think is the most reasonable explanation for the Hurst phenomenon? Base your answer upon references given in this chapter and elsewhere.

10.4 Using equations, explain the basic mathematical design and main purposes of the shifting level models referred to in Section 10.3.3.

10.5 Mention three types of yearly time series which could be appropriately modelled by FGN models. Provide both physical and statistical justifications for your suggestions.

10.6 In Section 10.4.5, a procedure is given for claculating a one step ahead MMSE (minimum mean square error) forecast for a FGN model. Develop a formula for determining l step ahead MMSE forecasts for a FGN model where $l \geq 1$.

10.7 In Section 10.4.6, seven methods are presented for approximately simulating FGN. Select any two of these techniques and explain using equations why these methods do not exactly simulate FGN.

10.8 In Table 10.4.4, the AIC is employed to decide upon whether or not FGN or ARMA models should be used for modelling six annual riverflow time series. Carry out a similar type of study for six annual time series that are not average yearly riverflows. Comment upon the results.

10.9 Within your field of study, select a statistic which is of direct interest to you. For example, you may be a hydrologist who is interested in floods or droughts. Explain how you would carry out simulation experiments to determine whether or not time series models fitted to your data sets preserve the historical statistics that are important to you.

10.10 Carry out the simulation study that you designed in the previous question.

10.11 Summarize Tsay's (1992) approach for ascertaining whether a fitted model preserves important historical statistics. Compare Tsay's procedure to the one presented in Section 10.6.

10.12 Explain how the research of Klemes and Klemes (1988) sheds light on the Hurst phenomenon.

REFERENCES

DATA SETS

De Geer, G. (1940). *Geochronologia Suecica Principles.* Almqvist and Wiksells, Stockholm.

Kendall, M. G. and Stuart, A. (1963). *The Advanced Theory of Statistics, Vol. 1, Distribution Theory.* Hafner, New York.

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society,* 79:242-261.

Schulman, E. (1956). *Dendroclimatic Changes in Semi-Arid America.* University of Arizona Press, Tucson, Arizona.

Stokes, M. A., Drew, L. G. and Stockton, C. W. (1973). Tree ring chronologies of Western America. Chronology Series 1, Laboratory of Tree Ring Research, University of Arizona, Tucson, Arizona.

Toussoun, O. (1925). Memories sur l'histoire de Nil, memoires de l'institut d'Egypte. *Imprimierie de l'Institut Francais d'Archeologie Orientale,* Cairo, pages 8-10.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960.* Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

FRACTIONAL GAUSSIAN NOISE

Cox, D. R. (1984). Long range dependence: A review. In David, H. A. and David, H. T., Editors, *Proceedings of the 50th Anniversary Conference of the Iowa State Statistical Laboratory,* held in Ames, Iowa, June 13-15, 1983, pages 55-74. The Iowa State University Press.

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models. *Advances in Applied Probability,* 8:339-364.

Garcia, L. E., Dawdy, D. R., and Mejia, J. M. (1972). Long memory monthly streamflow simulation by a broken line model. *Water Resources Research,* 8(4):1100-1105.

Lettenmaier, D. P. and Burges, S. J. (1977). Operational assessment of hydrologic models of long-term persistence. *Water Resources Research,* 13(1):113-124.

Mandelbrot, B. B. (1965). Une classe de processus stochastiques homothetiques a soi: Application a la loi climatologique de H. E. Hurst. *Compt. Rend. Acad. Sci.,* 260:3274-3276.

Mandelbrot, B. B. (1971). A fast fractional Gaussian noise generator. *Water Resources Research,* 7(3):543-553.

Mandelbrot, B. B. (1972). Broken line process derived as an approximation to fractional noise. *Water Resources Research,* 8(5):1354-1356.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional Brownian motion, fractional noises and applications. *Soc. Ind. Appl. Math. Rev.,* 10(4):422-437.

Mandelbrot, B. B. and Wallis, J. R. (1968). Noah, Joseph and operational hydrology. *Water Resources Research*, 4(5):909-918.

Mandelbrot, B. B. and Wallis, J. R. (1969a). Computer experiments with fractional Gaussian noises, 1, Averages and variances. *Water Resources Research*, 5(1):228-241.

Mandelbrot, B. B. and Wallis, J. R. (1969b). Computer experiments with fractional Gaussian noises, 2, Rescaled ranges and spectra. *Water Resources Research*, 5(1):242-259.

Mandelbrot, B. B. and Wallis, J. R. (1969c). Computer experiments with fractional Gaussian noises, 3, Mathematical appendix. *Water Resources Research*, 5(1):260-267.

Mandelbrot, B. B. and Wallis, J. R. (1969d). Some long-run properties of geophysical records. *Water Resources Research*, 5(2):321-340.

Mandelbrot, B. B. and Wallis, J. R. (1969e). Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. *Water Resources Research*, 5(5):967-988.

Matalas, N. C. and Wallis, J. R. (1971). Statistical properties of multivariate fractional noise processes. *Water Resources Research*, 7(6):1460-1468.

McLeod, A. I. and Hipel, K. W. (1978b). Comment on modelling monthly hydrologic persistence by G. K. Young and R. U. Jettmar. *Water Resources Research*, 14(4):699-702.

Mejia, J. M., Rodriquez-Iturbe, I. and Dawdy, D. R. (1972). Streamflow simulation, 2, The broken line model as a potential model for hydrologic simulation. *Water Resources Research*, 8(4):931-941.

Noakes, D. J., Hipel, K. W., McLeod, A. I., Jimenez, J. and Yakowitz, S. (1988). Forecasting annual geophysical time series. *International Journal of Forecasting*, 4:103-115.

O'Connell, P. E. (1974a). Stochastic modelling of long-term persistence in stream flow sequences. Ph.D. thesis, Civil Engineering Department, Imperial College, London, England.

O'Connell, P. E. (1974b). A simple stochastic modelling of Hurst's law. In *Proceedings of the International Symposium on Mathematical Models in Hydrology*. IAHS publication, No. 100.

Rodriguez-Iturbe, I., Mejia, J. M. and Dawdy, D. R. (1972). Streamflow simulation, 1, a new look at Markovian models, fractional Gaussian noise and crossing theory. *Water Resources Research*, 8(4):921-930.

Scheidegger, A. E. (1970). Stochastic models in hydrology. *Water Resources Research*, 6(3):750-755.

Taqqu, M. S. (1979). Self-similar processes and related ultraviolet and infrared castrophes. Technical Report 423, School of Operations Research and Industrial Engineering, Cornell University.

Young, G. K. and Jettmar, R. U. (1976). Modelling monthly hydrologic persistence. *Water Resources Research*, 12(5):829-835.

HURST PHENOMENON

Anis, A. A. (1955). The variance of the maximum of partial sums of a finite number of independent normal variates. *Biometrika*, 42:96-101.

Anis, A. A. (1956). On the moments of the maximum of the partial sums of a finite number of independent normal variates. *Biometrika*, 43:70-84.

Anis, A. A. and Lloyd, E. H. (1953). On the range of partial sums of a finite number of independent normal variates. *Biometrika*, 40:35-42.

Anis, A. A. and Lloyd, E. H. (1975). Skew inputs and the Hurst effect. *Journal of Hydrology*, 26:39-53.

Anis, A. A. and Lloyd, E. H. (1976). The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika*, 63:111-116.

Barnard, G. A. (1956). Discussion on methods of using long-term storage in reservoirs. *Proceedings of the Institute of Civil Engineers*, 1:552-553.

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science* 7(4):404-427.

Berman, S. M. (1964). Limiting distributions of the maximum of a diffusion process. *Annals of Mathematical Statistics*, 35:319-329.

Bhattacharya, R. N., Bupta, V. K. and Waymire, E. C. (1983). The Hurst effect under trends. *Journal of Applied Probability* 20(3):649-662.

Boes, D. C. and Salas-La Cruz, J. D. (1973). On the expected range and expected adjusted range of partial sums of exchangeable random variables. *Journal of Applied Probability*, 10:671-677.

Feller, W. (1951). The asymptotic distribution of the range of sums of independent random variables. *Annals of Mathematical Statistics*, 22:427-432.

Gomide, F. L. S. (1975). Range and deficit analysis using Markov chains. Hydrology Paper no. 79, Colorado State University, Fort Collins, Colorado.

Gomide, F. L. S. (1978). Markovian inputs and the Hurst phenomenon. *Journal of Hydrology*, 37:23-45.

Granger, C. W. J. and Orr, D. (1972). Infinite variance and research strategy in time series analysis. *Journal of the American Statistical Association*, 67(338):275-285.

Hipel, K. W. (1975). Contemporary Box-Jenkins Modelling in Water Resources. PhD thesis, University of Waterloo, Waterloo, Ontario.

Hipel, K. W. and McLeod, A. I. (1978a). Preservation of the rescaled adjusted range, 2, simulation studies using Box-Jenkins models. *Water Resources Research*, 14(3):509-516.

Hipel, K. W. and McLeod, A. I. (1978b). Preservation of the rescaled adjusted range, 3, Fractional Gaussian noise algorithms. *Water Resources Research*, 14(3):517-518.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770-808.

Hurst, H. E. (1956). Methods of using long-term storage in reservoirs. *Proceedings of the Institute of Civil Engineers*, 1:519-543.

Klemes, V. (1974). The Hurst phenomenon a puzzle? *Water Resources Research* 10(4):675-688.

Klemes, V. and Klemes, I. (1988). Cycles in finite samples and cumulative processes of higher orders. *Water Resources Research*, 24(1):93-104.

Klemes, V., Srikanthan, and McMahon, T. A. (1981). Long-memory flow models in reservoir analysis: What is their practical value? *Water Resources Research*, 17(3):737-751.

Kunsch, H. (1986). Discrimination between monotonic trends and long-range dependence. *Journal of Applied Probability*, 23:1025-1030.

Matalas, N. C. and Huzzen, C. S. (1967). A property of the range of partial sums. Paper presented at the International Hydrology Symposium, Colorado State University, Fort Collins, Colorado.

McLeod, A. I. and Hipel, K. W. (1978a). Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon. *Water Resources Research*, 14(3):491-508.

Moran, P. A. P. (1959). *The Theory of Storage*. Methuen, London.

Moran, P. A. P. (1964). On the range of cumulative sums. *Annals of the Institute of Statistical Mathematics*, 16:109-112.

O'Connell, P. E. (1976). Skew inputs and the Hurst effect - A comment. *Journal of Hydrology*, 31:185-191.

Poreda, G. and Mesa, O. J. (1988). Acera de la existencia del fen´omena de Hurst. *Primer Seminario Latinoamericano Sobre Aprovechamiento de Recursos Hidr´aulicos*, Universidad Nacional, Medellin, Colombia, 32 pp.

Salas, J. D., Boes, D. C., Yevjevich, V., and Pegram, G. G. S. (1979). Hurst phenomenon as a pre-asymptotic behaviour. *Journal of Hydrology*, 44:1-15.

Salas-La Cruz, J. D. and Boes, D. C. (1974). Expected range and adjusted range of hydrologic sequences. *Water Resources Research*, 10(3):457-463.

Siddiqui, M. M. (1976). The asymptotic distribution of the range and other functions of partial sums of stationary processes. *Water Resources Research*, 12(6):1271-1276.

Sim, C. H. (1987). Model for river flow time series. *Water Resources Research* 23(1):32-36.

Solari, M. E. and Anis, A. A. (1957). The mean and variance of the maximum of the adjusted partial sums of a finite number of independent normal variates. *Annals of Mathematical Statistics*, 28:706-716.

Taqqu, M. (1970). Note on evaluation of R/S for fractional noises and geophysical records. *Water Resources Research*, 6(1):349-350.

Wallis, J. R. and Matalas, N. C. (1970). Small sample properties of H and K estimators of the Hurst coefficient h. *Water Resources Research*, 6(6):1583-1594.

Wallis, J. R. and O'Connell, P. E. (1973). Firm reservoir yield - how reliable are hydrological records. *Hydrological Sciences Bulletin*, 39:347-365.

SHIFTING LEVEL MODELS

Ballerini, R. and Boes, D. C. (1985). Hurst behavior of shifting level processes. *Water Resources Research*, 12(11):1642-1648.

Boes, D. C. and Salas, J. D. (1978). Nonstationarity of the mean and the Hurst phenomenon. *Water Resources Research,* 14(1):135-143.

D'Astous, F., Hipel, K. W. and McLeod, A. I. (1979). Comment on evidence for nonstationarity as a physical explanation of the Hurst phenomenon by K.W. Potter. *Water Resources Research,* 15(2):501-504.

Hurst, H. E. (1957). A suggested statistical model of some time series which occur in nature. *Nature,* 180:494.

Klemes, V. (1974). The Hurst phenomenon: A puzzle? *Water Resources Research,* 10(4):675-688.

Potter, K. W. (1976). Evidence for nonstationarity as a physical explanation of the Hurst phenomenon. *Water Resources Research,* 12(5):1047-1052.

Salas, J. D. and Boes, D. C. (1980). Shifting level modelling of hydrologic series. *Advances in Water Resources,* 3:59-63.

Smith, J. A. (1988). A model of daily municipal water use for short-term forecasting. *Water Resources Research* 24(2):201-206.

STATISTICS

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Fisher, R. A. (1970). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburg, England.

Gnedenko, B. V. (1968). *Theory of Probability.* Chelsea, New York.

Healy, M. J. R. (1968). Algorithm AS6, triangular decomposition of a symmetric matrix. *Journal of the Royal Statistical Society, Series C (Applied Statistics),* 17:195-197.

Knuth, D. F. (1973). *The Art of Programming, Vol. 3.* Addison-Wesley, Reading, Massachusetts.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* John Wiley, New York, second edition.

TIME SERIES ANALYSIS

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* AC19:716-723.

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

Cox, D. R. (1991). Long-range dependence, non-linearity and time irreversibility. *Journal of Time Series Analysis* 12(4):329-335.

Lawrance, A. J. and Kottegoda, N. T. (1977). Stochastic modelling of riverflow time series. *Journal of the Royal Statistical Society, Series A,* 140(Part 1):1-47.

McLeod, A. I. and Hipel, K. W. (1978c). Simulation procedures for Box-Jenkins models. *Water Resources Research* 14(5):969-975.

Tsay, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *Applied Statistics* 41(1):1-15.

CHAPTER 11

FRACTIONAL AUTOREGRESSIVE-MOVING AVERAGE MODELS

11.1 INTRODUCTION

As explained in detail in Chapter 10, the well known *Hurst Phenomenon* defined in Section 10.3.1 stimulated extensive research in the field of stochastic hydrology. One valuable by-product of this research was the development of *long memory models* (see Sections 2.5.3, 10.3.3 and 11.2.1 for a definition of long memory). In particular, the *fractional Gaussian noise (FGN)* model of Section 10.4 possesses long memory and was developed within stochastic hydrology as an attempt to explain the Hurst Phenomenon through the concept of *long term persistence.*

The FGN model is not the only kind of stochastic model having long memory. As a matter of fact, due to its rather inflexible design and the difficulties encountered when applying it to real data (see Section 10.4), researchers have studied a variety of long memory models. The objective of this chapter is to present the most flexible and useful class of long memory models that have currently been developed. More specifically, this family is called the *fractional autoregressive-moving average (FARMA) group of models* (Hosking, 1981; Granger and Joyeux, 1980) because it arises as a natural extension of the ARIMA(p,d,q) models of Chapter 4. By allowing the parameter d in an ARIMA(p,d,q) model to take on real values, the resulting FARMA model possesses long memory for d falling within the range $0 < d < 1/2$.

A sound explanation for the Hurst phenomenon is presented in Section 10.6. In particular, by properly fitting ARMA models to a variety of geophysical time series, it is shown using simulation that the *ARMA models statistically preserve the Hurst statistics* consisting of the *rescaled adjusted range (RAR)* and the *Hurst coefficient* K. Because FARMA models are simply extensions or generalizations of ARMA (Chapter 3) and ARIMA (Chapter 4) models, one could also consider FARMA models in statistical experiments similar to those given in Section 10.6. Nonetheless, from a physical viewpoint hydrologic phenomena such as annual riverflows do not possess long memory or persistence since current flows do not depend upon annual flows that took place hundreds or thousands of years ago. Hence, for these kinds of series, ARMA models can adequately explain the Hurst phenomenon. However, the reader should keep in mind that there may be series that have long term memory and for these data one can employ FARMA models.

In the next section, the FARMA model is defined and some of its main statistical properties are described. Within Section 11.3, it is explained how FARMA models can be fitted to time series by following the identification, estimation and diagnostic check stages of *model construction.* Although good model building tools are now available, further research is required for developing more comprehensive estimation procedures. Methods for *simulating* and *forecasting* with FARMA models are given in Section 11.4. Before the conclusions, FARMA models are fitted to hydrological time series to illustrate how they are applied in practice. Parts of the presentations provided in Sections 11.2 to 11.5 were originally given in a paper by Jimenez et al. (1990).

11.2 DEFINITIONS AND STATISTICAL PROPERTIES

11.2.1 Long Memory

Persistence or *long term memory* is the term used to describe a time series that has either an autocorrelation structure that decays to zero slowly with increasing lag or equivalently a spectral density that is highly concentrated at frequencies close to zero. This autocorrelation structure suggests that the present state of the process must be highly dependent on values of the time series lying far away in the past, and, hence, to model the process the whole past should be incorporated into the description of the process.

A variety of precise mathematical definitions for long memory are given by authors such as Eberlein and Taqqu (1986), Davison and Cox (1989) as well as other authors cited in this chapter and Chapter 10. A simple definition that captures the essence of persistence, is the one presented in Sections 2.5 and 10.3.3. More specifically, a time series process can be classified according to the behaviour of the memory of the process where memory is defined as

$$M = \sum_{k=-\infty}^{\infty} |\rho_k|, \qquad [11.2.1]$$

where ρ_k is the theoretical ACF at lag k for the process. A long term memory process is defined as a process with $M = \infty$, whereas a short term memory process has $M < \infty$. The M term is often used as a mixing coefficient for stationary time series (Brillinger, 1975) to indicate the rate at which the present values of the time series are independent of the far away past values. The asymptotic independence between values of the time series well spaced in time where the mixing rate is given by $M < \infty$, has been traditionally used by time series analysts to prove results relating to normality, asymptotic behaviour of a quantity like the sample ACF, parameter estimates obtained either by maximum likelihood estimation or by the method of moments, hypothesis testing, and Portmanteau tests. Hence, most of the findings usually used in time series analysis are not necessarily true for long term memory processes because these processes have an infinite memory.

Besides hydrology, meteorology and geophysics, the classification of time series according to short and long memory may be useful in other areas (Cox, 1984; Parzen, 1982) such as economics (Granger, 1980; Granger and Joyeux, 1980). This classification has been used even with other types of stochastic processes (Cox, 1984), although the memory has had other definitions. An alternative definition of long term memory, essentially equivalent to the definition given above, is to consider time series processes whose ACF decays as

$$\rho_k = O(k^{-a}), \qquad [11.2.2]$$

where a lies in the interval (0,1).

An advantage of the FARMA family of models, defined in the next subsection, is that it can describe both short and long term memory. Furthermore, it constitutes a direct generalization of the ARMA and ARIMA models of Chapters 3 and 4, respectively.

11.2.2 Definition of FARMA Models

As explained in Chapter 4, a device frequently used in time series modelling is differencing the series, if it is thought that its mean function is time dependent. A time dependent mean could produce sample autocorrelations that, rather than decaying to zero exponentially like the ACF's of ARMA models, decay to zero much more slowly. In fact, if the rate of decay of the ACF seems to depend linearly upon the lag the usual approach is to work with the first differences of the time series. For the type of processes studied in this chapter, the ACF decays to zero at a rate slower than exponential, but faster than linear. This suggests the use of a device similar to the usual differencing operator, to model time series having a slowly decaying ACF with long memory. In fact, FARMA models generalize in a natural form the concept of ARIMA time series models containing differencing operators.

The FARMA family of models is a generalization of the ARIMA models of Chapter 4 which in turn constitute an extension of the ARMA models of Chapter 3. To define FARMA models, the concept of differencing is generalized by means of the filter

$$\begin{aligned}\nabla^d(B) = (1-B)^d &= \sum_{j=0}^{\infty}\begin{pmatrix}d\\k\end{pmatrix}(-B)^k \\ &= 1 - dB - \frac{1}{2}d(1-d)B^2 - \frac{1}{6}d(1-d)(2-d)B^3 - \cdots \end{aligned} \qquad [11.2.3]$$

where B is the backward shift operator. For an ARIMA model, the values of d in the filter in [11.2.3] are restricted to be zero when the series being modelled is stationary and to be a positive integer when the series must be differenced to remove nonstationarity. When d can be *fractional,* and hence take on real values, the above filter becomes the one used with FARMA models. As is explained in Section 11.2.3 on the statistical properties of FARMA models, the value of d controls the memory of the process.

As originally suggested independently by Hosking (1981) and Granger and Joyeux (1980), a *FARMA(p,d,q) model* for modelling a series z_t is defined as

$$\phi(B)\nabla^d z_t = \theta(B)a_t \qquad [11.2.4]$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the autoregressive (AR) operator of order p having the AR parameters $\phi_1, \phi_2, \ldots, \phi_p$; $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the moving average (MA) operator of order q having the MA parameters $\theta_1, \theta_2, \ldots, \theta_q$; ∇^d is the fractional differencing operator defined in [11.2.3]; and a_t is a white noise process that is identically and independently distributed with a mean of zero and variance of σ_a^2 (i.e. IID$(0,\sigma_a^2)$). As is also the case for the standard ARMA model of Chapter 3, the operators $\phi(B)$ and $\theta(B)$ are assumed to have all roots lying outside the unit circle and to have no common roots. Finally, no mean level, μ, is written in [11.2.4] since $\nabla^d \mu = 0$ for positive d.

One can write the FARMA process in [11.2.4] as

$$\nabla^d z_t = \frac{\theta(B)}{\phi(B)}a_t \qquad [11.2.5]$$

One can interpret the short memory component of the FARMA process as being modelled by applying the usual ARMA filter given by $\theta(B)/\phi(B)$ to the a_t time series that is IID$(0,\sigma_a^2)$. The *fractional differencing filter* ∇^d handles the long memory part of the overall process.

For a nonseasonal FARMA model, the notation FARMA(p,d,q) is employed where p and q are the orders of the AR and MA operators, respectively, and d is a parameter in the filter in [11.2.3] and can take on real values. When d is a positive integer, the FARMA(p,d,q) model is equivalent to an ARIMA(p,d,q) model where the acronym ARIMA stands for autoregressive integrated moving average (see Chapter 4). If $d = 0$, the FARMA(p,d,q) model is identical to a short memory ARMA(p,q) model of Chapter 3. When $p = q = 0$, the FARMA(p,d,q) model reduces to

$$\nabla^d z_t = a_t \qquad [11.2.6]$$

which is called a *fractional differencing model*. The labels that can be used for the various types of FARMA, ARIMA and ARMA models are listed in Table 11.2.1. In this table, the FARMA(p,d,q) model is the most general and it contains all the other models as subsets.

As an example of how to write a specific FARMA(p,d,q) model, consider the case of a FARMA(0,0.3,1) for which $p = 0$, $q = 1$ and d has a real value of 0.3. From [11.2.4] this model is given as

$$(1 - B)^{0.3} z_t = (1 - \theta_1 B)a_t$$

Using [11.2.3], the fractional differencing operator is expanded as

$$(1 - B)^{0.3} = 1 - 0.3B - \frac{1}{2}(0.3)(1 - 0.3)B^2 - \frac{1}{6}0.3(1 - 0.3)(2 - 0.3)B^3 - \cdots$$

$$= 1 - 0.3B - 0.105B^2 - 0.060B^3 - \cdots$$

Substituting the expanded fractional differencing operator into the equation for the FARMA(0,0.3,1) model results in

$$(1 - 0.3B - 0.105B^2 - 0.059B^3 - \cdots)z_t = (1 - \theta_1 B)a_t$$

or

$$z_t - 0.3z_{t-1} - 0.105z_{t-2} - 0.059z_{t-3} - \cdots = (1 - \theta_1 B)a_t$$

or

$$z_t = 0.3z_{t-1} + 0.105z_{t-2} + 0.059z_{t-3} + \cdots + a_t - \theta_1 a_{t-1} \qquad [11.2.7]$$

From this equation, one can see that the weights for the z_t terms are decreasing as one goes further into the past.

For the theoretical definition of the FARMA(p,d,q) model in [11.2.4], the a_t series is assumed to be IID($0,\sigma_a^2$). In order to develop estimation and other model construction methods, usually the a_t's are assumed to be normally distributed. Recall that the assumption that the a_t's are NID($0,\sigma_a^2$) for application purposes is also invoked in Part III for the ARMA and ARIMA models of Chapters 3 and 4, respectively, as well as most other models presented in this book.

Table 11.2.1. Names of models.

Values of d	Values of p	Values of q	Equivalent Model Names	Chapters
Real value	p	q	FARMA(p,d,q)	11
Real value	0	0	FARMA(0,d,0), Fractional Differencing	11
Positive Integer	p	q	ARIMA(p,d,q), FARMA(p,d,q) for which d is a positive integer	4
Positive Integer	0	q	ARIMA(0,d,q), IMA(d,q), FARMA(0,d,q) for which d is a positive integer	4
0	p	q	ARMA(p,q), FARMA(p,0,q), ARIMA(p,0,q)	3
0	p	0	ARMA(p,0), AR(p), FARMA(p,0,0), ARIMA(p,0,0)	3
0	0	q	ARMA(0,q), MA(q), FARMA(0,0,q), ARIMA(0,0,q)	3
0	0	0	ARMA(0,0), FARMA(0,0,0), ARIMA(0,0,0), White Noise	3

When the residuals of a fitted FARMA model and, hence, the original series are not normally distributed, one approach to overcome this problem is to invoke the Box-Cox transformation defined in [3.4.30]. Subsequent to this, one can estimate the parameters of all the model parameters, including d, for the FARMA model fitted to the transformed series.

Three classes of seasonal models are given in Part VI of the book. The definition for nonseasonal FARMA models can be easily extended to create *long memory seasonal FARMA models* for each of the three kinds of seasonal models. To create a seasonal FARMA model, which is similar to the seasonal ARIMA model of Chapter 12, one simply has to incorporate a seasonal fractional differencing operator as well as seasonal AR and MA operators into the basic nonseasonal FARMA model in [11.2.4]. A deseasonalized FARMA model is formed by fitting a nonseasonal FARMA model to a series which has been first deseasonalized using an appropriate deseasonalization technique from Chapter 13. To obtain a periodic FARMA model that reflects the periodic ARMA model of Chapter 14, one simply defines a separate nonseasonal FARMA model for each season of the year. Future research could concentrate on developing comprehensive model building techniques, especially for the cases of seasonal FARMA and periodic FARMA models. Hui and Li (1988) have developed maximum likelihood estimators for use with periodic FARMA(0,d,0) (i.e. periodic fractional differencing) and periodic FARMA(p,d,0) models.

Another function for FARMA modelling is to allow the noise terms of transfer function-noise (Part VII), intervention (Part VIII), and multivariate ARMA (Part IX) models to follow a FARMA model. The definitions of these models are simple. However, the development of model construction techniques, especially efficient estimation methods, would be a formidable task. Hence, this should only be undertaken if practical applications using real world data indicate a need for these kinds of long memory models.

Keeping in mind that FARMA modelling can be expanded in many directions, the rest of this chapter is restricted to the case of nonseasonal FARMA models. In the next subsection, some theoretical properties of the FARMA(p,d,q) model in [11.2.4] are given.

11.2.3 Statistical Properties of FARMA Models

As explained by Hosking (1981), the FGN model of Section 10.4 is in fact a discrete-time analogue of continuous-time fractional noise. Another discrete time version of continuous-time fractional noise is the fractional differencing (i.e. FARMA(0,d,0)) model in [11.2.6]. An advantage of the fractional differencing model over FGN is that it can be expanded to become the comprehensive FARMA(p,d,q) model in [11.2.4], which in turn is a generalization of the ARIMA model.

The basic properties of FARMA processes are presented by Hosking (1981) and Granger and Joyeux (1980). As explained by Jimenez et al. (1990), they found among other things that:

(a) For the process to be stationary, $d < 0.5$ and all the roots of the characteristic equation $\phi(B) = 0$ must lie outside the unit circle.

(b) For the process to be invertible, $d > -0.5$ and all the roots of the characteristic equation $\theta(B) = 0$ must lie outside the unit circle.

(c) Because of (a) and (b), if $-\frac{1}{2} < d < \frac{1}{2}$, the FARMA($p$,$d$,$q$) process is both stationary and invertible.

(d) For $0 < d < \frac{1}{2}$, the process has long memory (see Section 11.2.1 for definitions of long memory).

(e) The ACF behaves as

$$\rho_k = 0(k^{-1+2d}). \qquad [11.2.8]$$

(f) The process is self-similar, which means that the stochastic properties of the process are invariant under changes of scale.

Several probabilists (Rosenblatt, 1961, 1979, 1981; Taqqu, 1975) have studied the behaviour of statistics derived from time series processes where the ACF behaves as in [11.2.8] and d is positive. They found that:

(g) The sample mean times $N^{1/2+d}$, where N is the number of observations, converges in law to a normal random variable.

(h) The sample autocovariances do not converge asymptotically to a normal random variable.

The result in (g) about the mean is of some interest to hydrologists because it has been found that processes thought to possess long term memory have a sample mean that seems to indicate slow changes in trend. This wandering of the sample mean can be explained in terms of (g) above. This shifting level process for modelling a changing mean is referred to in Section 10.3.3 and references are provided at the end of Chapter 10. Arguments to show that persistence in geophysical processes is due to a slowly changing trend cannot be based just on statistical behaviour but should use geophysical insight. Note also that the above results are not restricted to the FARMA process case but that they are valid for any time series whose ACF behaves as in [11.2.8].

An important, although seemingly trivial extension of the original definition of FARMA processes by Hosking (1981) and Granger and Joyeux (1980) is to relax the assumption that the mean of the time series is zero. The extension of the model given above to the case of a nonzero mean is straightforward. However, what is very important to note is that if a constant, in particular the mean, is passed through the filter ∇^d, the output, for the case of a positive d, is zero. Hence, the mean of the process does not have to appear in the equations that define the model. Nevertheless, it should be noted that the mean is a well defined quantity for this process when $d < 0.5$.

The aforementioned property is very important for determining the stochastic properties of the estimates for the parameters. This is because the sample mean can be used as an estimate for the mean of the time series and the slow rate of convergence of the sample mean as given in (g) above does not affect the asymptotic rate of convergence of the estimates for the other parameters to a Gaussian random variable, where this rate is the usual N^{-1}.

Another interesting feature is that the filter ∇^d can smooth some special trends as can be seen easily for the case $d = 1$ when the trend is a straight line. When $0 \leq d < 0.5$ the filter ∇^d smooths slowly changing trends. Hence, even if the process mean is slowly changing, FARMA models could be used to model the time series in much the same way that ARIMA models are employed with a deterministic drift component.

Another consequence of the fact that

$$\nabla^d(z_t - \mu) = \nabla^d z_t, \quad d > 0 \qquad [11.2.9]$$

where z_t is the value of the process at time t with a theoretical mean μ, is that the process behaviour is independent of the mean. In the stationary ARMA process, on the other hand, the local behaviour of the process does depend on the mean. This can be seen by considering the value of the process conditioned on the past as given by $E\{z_{t+1}|z_s, s \leq t\}$. In the ARMA case, this quantity depends on $\mu = E\{z_t\}$ but in the FARMA case with $d > 0$, it does not. In the remaining parts of this section, unless stated to the contrary, the mean μ will be assumed equal to zero.

An important consequence of the slow rate of decay to zero of the ACF as given by [11.2.8] is that Bartlett's formula (Bartlett, 1946) for the variances and the covariances of the estimated autocovariance function (ACVF), $\{\hat{\gamma}_k\}$, has to be modified accordingly. In fact, the exact formula for the variance is given by

$$var(\hat{\gamma}_k) = N^{-1} \sum_{m=-(N-k)+k}^{(N-k-1)} \left\{1 - |m| + \frac{k}{N}\right\}\left\{\gamma_m^2 + \gamma_{m+k}\gamma_{m-k}\right\} \qquad [11.2.10]$$

Then, by [11.2.8]

$$var(\hat{\gamma}_k) = \begin{cases} 0(N^{-1}), & \text{if } d \leq 0.25 \\ 0(N^{4d-2}), & \text{if } d > 0.25 \end{cases} \qquad [11.2.11]$$

Hence, if $d < 0.25$ then $var(\hat{\gamma}_k) = 0(N^{-1})$, which is the same order as in the case of a short memory process. However, if $0.25 < d < 0.5$ the order of $var(\hat{\gamma}_k)$ is larger than N^{-1}. In fact as d approaches 0.5 the variance approaches a quantity of order one. This implies that the stochastic variability of the estimated ACVF is higher for long term memory processes with $0.25 < d < 0.5$ than for short term memory processes. Moreover, the order of the variance depends on the unknown quantity d. Finally, similar results are valid for the covariances of the estimated ACF.

An interesting subset of the FARMA(p,d,q) family of processes in [11.2.4] is the FARMA$(0,d,0)$ process in [11.2.6] which is referred to as the fractional differencing model. This model has been studied in some detail and expressions for the ACF, partial autocorrelations function (PACF), partial linear regression coefficients, and inverse autocorrelations are known (Hosking, 1981, 1984, 1985). One important fact about the stochastic behaviour of a FARMA$(0,d,0)$ process is that all its autocorrelations are positive if d is positive, and they are negative otherwise. Also, all the partial autocorrelations of the FARMA$(0,d,0)$ model have the same sign as the *persistence parameter d*, and their rate of decay to zero is of the same order as the inverse of the lag. Because of these limitations of the structure of the ACF, fractionally differenced noise is passed through an ARMA filter in order to obtain a richer autocorrelation structure within the framework of a FARMA(p,d,q) process.

As suggested by Jiminez et al. (1990), it is possible to generalize the filter $(1-B)$ in another form, which is closely related to the $(1-B)^d$ filter in [11.2.3]. In particular, this filter is defined by $(1+B)^d$. Note that the associated transfer function also has a root on the unit circle at $B = -1$. The coefficients of this filter are the same as those of the filter $(1-B)^d$ except for the sign and hence the process also has long term memory if $d > 0$, it is stationary if $d < 0.5$, and invertible if $d > -0.5$. However, the interesting fact is that although the absolute values of the autocorrelations are the same for both filters, the autocorrelations of the filter $(1+B)^d$ alternate in sign. More general autocorrelations structures could be obtained by generalizing the filters to accommodate complex roots on the unit circle. The class of processes studied in this chapter are particular cases of the more general processes that result by filtering white noise through the filters defined by $(1-\varepsilon B)^d$, where the parameter ε lies in the range $|\varepsilon| \leq 1$. In this chapter it is assumed that $\varepsilon = 1$, or -1.

11.3 CONSTRUCTING FARMA MODELS

11.3.1 Overview

To fit a FARMA(p,d,q) model to a given time series, one can follow the usual identification, estimation and diagnostic check stages of model construction. Model building procedures are fairly well developed for the case of the fractional differencing (i.e. FARMA$(0,d,0)$ model in [11.2.6]). However, further research is required for obtaining a comprehensive set of tools for building the FARMA(p,d,q) models in [11.2.4]. Of particular importance is the need for good estimation techniques that are both computationally and statistically efficient, as well as capable of estimating the mean level along with the other FARMA model parameters. Unlike ARIMA(p,d,q) models where d is fixed at zero or some positive integer value prior to estimating the other model parameters for the differenced series, one must, of course, estimate d in the FARMA(p,d,q) model simultaneously with the othcr model parameters.

11.3.2 Identification

To identify a suitable ARMA model (Chapter 3) or ARIMA model (Chapter 4) to fit to a given time series, one can examine the characteristics of graphs of the sample ACF, PACF, IACF and IPACF (Chapter 5). By knowing the behaviour of the theoretical ACF, PACF, IACF and IPACF for ARMA or ARIMA models, one can determine from the sample plots which parameters to include in the model. If more than one model is fitted to the series, an automatic selection criterion such as the AIC (see Section 6.3) can be used to select the best one.

The sample ACF, PACF, IACF and IPACF can also be used to identify a FARMA(p,d,q) model for fitting to a series. If the series is stationary and the sample ACF dies off slowly, then d should be estimated to account for this long term persistence. Hosking (1981) gives formulae for the theoretical ACF, PACF and IACF for the case of the fractional differencing model in [11.2.6]. Further research is required to obtain formulae for the theoretical PACF, IACF and IPACF for FARMA(p,d,q) models. By comparing the behaviour of the sample graphs to the theoretical findings one can decide upon which parameters to include in the FARMA(p,d,q) model. Additional procedures for model identification are presented in Section 11.5 with the applications.

11.3.3 Estimation

This section follows the research findings of Jimenez et al. (1990). However, the reader may also wish to refer to the FARMA estimation procedures presented by Boes et al. (1989), and by Brockwell and Davis (1987, pp. 464-478), as well. As noted earlier in Section 11.2.3, the coefficients of the filter $(1-B)^d$ and $(1+B)^d$ only differ in sign. Because the estimation results of this section are valid for both filters, everything is described only for the filter $\nabla^d=(1-B)^d$.

There are several estimation procedures available in the literature. Frequency domain methods do not seem to be as efficient as estimators based on the time domain representation. Hence, only time domain methods are considered here.

Because of the slow rate of convergence of the sample mean to the true mean as can be seen in (g) in Section 11.2.3, it is of utmost importance to find a more efficient estimator of the mean. The most obvious candidate is the maximum likelihood estimate of the mean (McLeod and Hipel, 1978a), which is given by

$$\hat{\mu} = (\mathbf{z}^T \Sigma^{-1} \mathbf{1})(\mathbf{1}^T \Sigma^{-1} \mathbf{1}) \qquad [11.3.1]$$

where $z^T = (z_1, z_2, \ldots, z_N)$ is the $1 \times N$ vector of observations, Σ is the autocorrelation matrix of the time series, and 1 represents a column vector of ones. However, it can be shown that the sample mean is efficient for the case when the persistence parameter d is nonnegative, and it is not efficient when the persistence parameter is negative. This agrees with the common knowledge that overdifferencing can lead to inefficient estimates. Although it is difficult to give a physical meaning to antipersistence, a negative value of d can be useful from a purely fitting point of view as it has been observed that sometimes FARMA models with negative d arise while fitting them to a time series, and, therefore, it is important in these cases to estimate the mean of the process using the maximum likelihood estimate as given by the above formula. The evaluation of the above formula can be performed efficiently using either Cholesky decomposition (Healy, 1968) of the inverse of Σ given by the partial linear regression coefficients, $\{\phi_{i,j}\}$ (which can be obtained easily by the Levison-Durbin algorithm (Durbin, 1960)), or by the Trench algorithm for the inverse of a Toeplitz matrix (Trench, 1964). For the particular case of fractionally differenced noise, Hosking (1981) gives a closed expression for the reflection coefficients or partial linear regression coefficients. Hence, in this case a closed expression for the maximum likelihood estimate of the mean is known. For the situation where $\varepsilon = -1$ in the filter, mentioned at the end of Section 11.2.3, this closed expression is still valid with appropriate sign changes. In terms of the partial linear regression coefficients, the following expression could be used to evaluate the maximum likelihood estimate $\hat{\mu}$

$$\hat{\mu} = N^{-1} \frac{\sum_{t=0}^{N-1} (z_t - \phi_{1,t} z_{t-1} - \phi_{2,t} z_{t-2} - \cdots - \phi_{t,t} z_o)}{\sum_{t=0}^{N-1} (1 - \phi_{1,t} - \phi_{2,t} - \cdots - \phi_{t,t})^2} \qquad [11.3.2]$$

In this section, it is assumed that the *persistence parameter d* is nonnegative, the sample mean is used as the estimate of the mean, and the sample mean has been subtracted from each observation.

There are two methods available to estimate the remaining parameters in the time domain: exact maximum likelihood estimation or an approximation of the filter ∇^d. Most of the maximum likelihood estimation algorithms depend on computing the one step ahead prediction errors, a_t, which can be computed in terms of the partial linear regression coefficients. These coefficients can be computed efficiently by the Durbin-Levinson algorithm. Finally, with these values of e_t the estimates of the parameters are obtained by minimizing the modified sum of squares function given by:

$$\ln l = \sum_{t=1}^{N} (N - t + 1) \ln(1 - \phi_{t,t}^2) + \sum_{t=1}^{N} a_t^2 / \sigma_t^2. \qquad [11.3.3]$$

Although the computation of estimates by maximum likelihood is statistically attractive, the amount of computations involved in the above scheme makes algorithms having fewer numbers of computations competitive alternatives.

The algorithm proposed by Li and McLeod (1986) is computationally economical and is presented in Appendix A11.1. The algorithm consists of approximating the filter ∇^d by the filter ∇_M^d where ∇_M^d is defined as the filter resulting by taking the first M terms of the filter ∇^d, i.e. by approximating the process by a "long" autoregression. Then the algorithm minimizes the sum of the squared residuals, where the residuals are obtained as the output of the filters ∇_M^d and the ARMA filter. To compute the residuals, an algorithm such as the one given by McLeod and Sales (1983) could be used. Also, as recommended by Box and Jenkins (1976) the sum of squared residuals could be extended back in time by backforecasting. Note that the approximation of ∇^d by ∇_M^d is not the optimal approximation in a least squares sense. However, since the order to M is comparable with N, it has to be very close to the optimal approximation. The order of approximation necessary to obtain consistence estimates has been found to be of the order of $N^{1/2}$ and an ad hoc rule is to fit time series with at least 50 observations. The order of truncation M is chosen as a number of between $N/4$ and $N^{1/2}$, by trying to balance the degree of approximation of the filter ∇_M^d to the filter ∇^d and the amount of computations involved. Nonetheless, for N close to 50, M is taken as half the number of observations. The amount of computations using this algorithm is much smaller than that for the maximum likelihood approach. Moreover, estimates obtained in this form are asymptotically equivalent to the maximum likelihood estimates and it seems that the finite sample estimates are generally close enough to the maximum likelihood estimates. Li and McLeod (1986) studied the asymptotical distributions of the estimates when the mean of the time series is known. They derived closed form expressions for the variances of the asymptotically normal distributions of the estimates. It can be demonstrated that the estimation of the mean by the sample mean does not affect the above asymptotic results. However, these results are not likely to hold for a finite sample size because of the long term persistence and the parameter d is constrained to lie in the open interval (-0.5, 0.5). In practice, the interval is closed and it can be observed using simulation that if the persistence parameter is close to 0.5, there is a high probability for the estimate of d to be equal to 0.5. A similar phenomenon was observed for the ARMA(0,1) model by Cryer and Ledolter (1981). Hence, the rate of convergence of the estimates depends on the parameters even for relatively large sample sizes of more than 200. Additionally, it should be noted that the above method is very similar to fitting an autoregressive process of order one if d is not close to 0.5, say less than 0.3.

Bootstrapping a Time Series Model

Because the FARMA model is an infinite autoregression and, moreover, is nonstationary when $d \geq 0.5$, it is expected that finite sample properties of the estimates are different than the large sample approximations. Consequently, it is interesting to obtain further information about these finite sample distributions. One interesting possibility to increase one's knowledge of the finite sample distribution of the estimates is by using the bootstrapping technique proposed by Cover and Unny (1986).

Since Efron (1979) proposed the bootstrap, there have been several proposals to extend the original technique to time series analysis. However, most of them have used a straightforward generalization of the original bootstrap with the consequence that what they did was to use distorted models. The idea of Cover and Unny (1986) is to inject randomness into the loss function by resampling the positions of the residuals and not the observations themselves (i.e. the time lags are resampled with replacement and with the same probability). This resampling of the time lags is interesting because of the nature of data that depends strongly on the time coordinates. Note also that unlike other resampling plans, the assumption that the fitted model is the true model is not crucial. Also, it can be applied to any time series model and not just to a FARMA model. Moreover, the idea is valid for other stochastic processes.

The technique can then be described as follows:

(a) Draw a random sample of size N with replacements from the integers between 1 and N;

(b) Obtain estimates of the parameters by minimizing the sum of the squared residuals a_t^2 with weights equal to the number of times that the number t appeared in the random sample in (a);

(c) Repeat (a) and (b) a large enough number of times to obtain reliable estimates of the distribution characteristics of the estimated parameters.

This technique can greatly increase one's information about the parameter estimates as can be seen in the applications. However, further theoretical results are required to confirm theoretically the finite sample validity of the bootstrap approach.

11.3.4 Diagnostic Checks

To ascertain if a calibrated FARMA(p,d,q) model adequately fits a given series, one can employ diagnostic checks similar to those given in Chapter 7 for ARMA and ARIMA models. The innovations of the FARMA model in [11.2.4] are assumed to be Gaussian, homoscedastic (i.e. have constant variance) and white. When, in practice, the residuals of the fitted model are not always normal/homoscedastic, this can often be overcome by transforming the data using the Box-Cox transformation of [3.4.30]. The parameters of the FARMA(p,d,q) model can then be estimated for the transformed series and the residuals once again subjected to diagnostic checks.

The most important innovation assumption is independence. If the residuals of the fitted model are correlated and not white, then a different FARMA model or, perhaps, some other type of model, should be fitted to the series. The best check for whiteness is to examine the residual autocorrelation function (RACF) for the calibrated model, as is also the case for ARMA and ARIMA models (see Chapter 7). The large-sample distribution of the RACF for a FARMA model is given by Li and McLeod (1986) who also present a modified Portmanteau test statistic to check for whiteness.

11.4 SIMULATION AND FORECASTING

11.4.1 Introduction

After a FARMA(p,d,q) model has been fitted to a given series, the calibrated model can be employed for applications such as simulation and forecasting. The purpose of this section is to present simulation and forecasting procedures for use with FARMA models. Techniques for simulating and forecasting with ARMA and ARIMA models are presented in Part IV of the book. Finally, forecasting experiments in which fractional differencing models are used, in addition to other kinds of models, are given in Section 8.3.

11.4.2 Simulating with FARMA Models

Based upon a knowledge of closed expressions for the partial linear regression coefficients, $\phi_{k,j}$, fast algorithms for generating synthetic sequences from FARMA models can be given. Partial linear regression coefficients are defined as the values of α_k that minimize

$$E\{z_t - \alpha_1 z_{t-1} - \cdots - \alpha_t z_0\}^2 \qquad [11.4.1]$$

where E is the expectation operator. Thus, they are the values that minimize the one step ahead forecast errors. As is well known, the time series process can be written in terms of the innovations as:

$$z_t = a_t + \phi_{1,t} z_{t-1} + \cdots + \phi_{t,t} z_0, \qquad [11.4.2]$$

where the innovations $\{a_t\}$ are a sequence of independent Gaussian random variables with mean 0 and variance $\sigma_t^2 = \prod_{j=1}^{t}(1 - \phi_{j,j}^2)$. First consider the case of simulating fractionally differenced noise. Expressions for $\phi_{k,t}$ are presented by Hosking (1981), and recursive expressions are given by:

$$\phi_{t,t} = d/(t-d)$$

$$\phi_{j,t} = \phi_{j+1,t}(j+1)(t-j-d)/((j-1-d)(t-j)), \qquad [11.4.3]$$

Consequently, to simulate a FARMA(0,d,0) noise model it is only necessary to compute recursively $\phi_{k,t}$, generate a normal random variable and then use [11.4.2].

To simulate using a FARMA(0,d,q) model, the fractionally differenced noise is generated and then passed through the moving average filter. When generating synthetic data using a FARMA(p,d,0) model, one possible approach is to simulate the FARMA(0,d,0) model using above algorithm and after choosing p initial values, which can be done using the method in McLeod and Hipel (1978b), as explained below, generate recursively the other simulated values. Finally, the general FARMA(p,d,q) case can be obtained by a combination of the above methods.

Another possible method to generate synthetic sequences (McLeod and Hipel, 1978a,b) is to obtain the Cholesky decomposition of the matrix of the theoretical autocorrelations, Σ, and to multiply this decomposition matrix by a vector of independent Gaussian variables with mean zero and desired variance (see Section 9.4 for the case of ARMA models). Finally, a mean correction is added to the series. However, although this method is attractive for other models, it may be less desirable than the method described above because it involves the computation of the matrix of autocorrelations and the theoretical autocorrelations are given in terms of hypergeometric functions (Hosking, 1981). Thus, the computation task time necessary to compute the autocorrelations is much bigger than the computation time necessary to pass FARMA(0,d,0) noise through the different filters. However, once the ACF has been calculated and the required Cholesky decomposition obtained, this method is useful if many independent realizations of the process are to be simulated. Finally, both methods are equivalent in the case of the FARMA(0,d,0) model.

11.4.3 Forecasting with FARMA Models

Forecasting by using ARMA models is generally most useful when the forecaster is just interested in one step ahead or two steps ahead forecasts. This is because the forecast functions produced by ARMA models converge exponentially fast to the mean of the time series. Hence, in ARMA models long term forecasts are given by the mean μ or some estimate of it. This is not the case when the process has a long term memory, as is clear from the definition of persistence. For the case of a long memory process, the forecasting functions still converge to the mean μ:

however, the rate of convergence is not exponential but slower. For persistent time series, the rate of decay of the forecast function depends on the degree of persistence that the process possesses.

Another consequence of persistence in forecasting is that the variance of the forecast function of a persistent process decays to the variance of the process, σ_z^2 at a rate that could be substantially slower than exponential, depending on the degree of persistence. Therefore, confidence bounds for the l-step ahead forecasts of persistent processes are smaller than those of short term memory processes, if l is bigger than two or three. This can be seen if the time series model is written as a linear process (Box and Jenkins, 1976)

$$z_t = \sum_{k=0}^{\infty} \alpha_k a_{t-k}. \qquad [11.4.4]$$

Then, the l-step ahead forecast, $\hat{z}_t[l]$, is given by

$$\hat{z}_t[l] = \sum_{k=0}^{\infty} \alpha_{k+l} a_{t-k} \qquad [11.4.5]$$

but, for a FARMA model, $\alpha_k \approx k^{-1-d}$. Therefore,

$$var\{\hat{z}_t[l]\} = \sigma_z^2 \approx 0(l^{-1+2d}) \qquad [11.4.6]$$

Equation [11.4.5] is most helpful for forecasting if estimates of a_t are available and if the coefficients α_k decay to zero fast enough so that the necessary truncation involved in the computation of $\hat{z}_t[l]$ as given in [11.4.5] produces a negligible error. However, for FARMA models these coefficients do not decay fast enough, and, hence, expressions for the forecast function $\hat{z}_t[l]$ that do not involve approximations could be useful. The method proposed is based on the AR form of the time series as given by [11.4.2]. The forecast function is given by

$$\hat{z}_t[l] = \phi_{1,t}(l) z_t + \phi_{2,t}(l) z_{t-1} + \cdots + \phi_{t,t}(l) z_0 \qquad [11.4.7]$$

where

$$\phi_{i,t}(l) = \phi_{l,t+l} + \sum_{j=1}^{l-1} \phi_{j,t+l-1} \phi_{i,t}(j). \qquad [11.4.8]$$

This expression has advantages over the formula given in [11.4.5] because it does not involve approximation either by truncation of an infinite series or in the computation of the residuals. Moreover, by using [11.4.7] it is possible to show that

$$\hat{z}_t[l] \approx \phi_{l,t+l} z_t + \cdots + \phi_{t+l,t+l} z_0. \qquad [11.4.9]$$

Hence, as discussed above the forecast function decays to the theoretical mean μ at a rate slower than exponential. For example, for the FARMA(0,d,0) model

$$\hat{z}_t(l) \approx \frac{l^{-d-1} z_t + \cdots + (t+l)^{-d-1} z_0}{(-d-1)!}. \qquad [11.4.10]$$

11.5 FITTING FARMA MODELS TO ANNUAL HYDROLOGICAL TIME SERIES

To demonstrate how FARMA models are applied in practice, FARMA models are fitted to the fourteen hydrological time series listed in Table 11.5.1. The data consists of eleven annual river flows in m^3 /s from different parts of the world, two records of average annual rainfall in mm, and an annual temperature series in degrees Celsius. Because efficient estimation procedures are available for use with FARMA(0,d,0) (i.e. fractional differencing) and FARMA(p,0,q) (i.e. ARMA(p,q)) models, these are the models which are considered for fitting to the series. Estimation methods for use with FARM(0,d,0) and ARMA(p,q) models are presented in Sections 11.3.3 and Chapter 5, respectively.

Table 11.5.1. Annual time series used in the applications of FARMA models.

	Descriptions	Geographical Locations	Time Spans	Lengths
(1)	Saugeen River	Walkerton, Ontario, Canada	1915-1976	62
(2)	Dal River	near Norslund, Sweden	1852-1922	70
(3)	Danube River	Orshava, Romania	1837-1957	120
(4)	French Broad River	Asheville, N. Carolina	1880-1900	70
(5)	Gota River	near Sjotop-Vannersburg, Sweden	1807-1957	150
(6)	McKenzie River	McKenzie Bridge, Oregon	1900-1956	56
(7)	Mississippi River	St. Louis, Missouri	1861-1957	96
(8)	Neumunas River	Smalininkai, Lithuania	1811-1943	132
(9)	Rhine River	Basle, Switzerland	1807-1957	150
(10)	St. Lawrence River	Ogdensburg, New York	1800-1930	131
(11)	Thames River	Teddington, England	1883-1954	71
(12)	Rainfall	Fortaleza, Brasil	1849-1979	131
(13)	Rainfall	Philadelphia	1800-1898	99
(14)	Average temperature	Central England	1723-1970	248

In practice, the definition of long term memory in terms of $M = \infty$ in [11.2.1] is difficult to check and instead the persistence criterion given in [11.2.8] is used. Hence, a sample ACF that decays slowly to zero could indicate that the time series has long term memory. For those time series whose sample ACF decays to zero at a hyperbolic rate, the possibility of modelling them by FARMA models is considered. Within the fourteen data sets, the St. Lawrence Riverflows and the Philadelphia Rainfall series show an estimated ACF that seems to decay to zero hyperbolically. Therefore, these two data sets present evidence that suggests the use of FARMA models to fit them. The graph of the St. Lawrence Riverflow series against time and its sample ACF are shown in Figures II.1 and 3.2.1, respectively. For other records such as the Saugeen Riverflows and Rainfall at Fortaleza, the evidence, as given by the estimated ACF's, in favour of a persistence parameter is not so strong but it is a possibility. However, it should be remarked that if the persistence parameter d is close to zero and, hence, between 0 and 0.2, detection of long term memory by visual inspection of the autocorrelations can be difficult. Moreover, because Bartlett's formula needs to be multiplied by a factor of order N^{-1+4d} ($d \geq 0.25$), for the case of a FARMA(p,d,q) process, visual inspection of the sample ACF should be used with care when it is suspected that the process under analysis could have long term memory.

If the process belongs to the FARMA family of models, the PACF should decay to zero at a hyperbolic rate. This rate is independent of the degree of persistence. However, for the case of a FARMA(0,d,0) process, long term memory implies that all the values of the PACF should be positive. This behaviour of the PACF for the FARMA(0,d,0) process suggests that to detect persistence, not only a hyperbolic decay of the PACF is of interest, but also the behaviour of the signs of the PACF. This suggests the use of a nonparametric sign test to test the signs of the estimated PACF. None of the estimated PACF's of the fourteen data sets show strong evidence of a hyperbolic rate of decay to zero. However, some of them like those for the St. Lawrence Riverflows and Philadelphia Rainfall series show PACF structures that are generally positive. A sign test of these PACF's gives further support for the conjecture that these time series demonstrate signs of persistence.

Another characteristic of a time series that could indicate the presence of persistence is the behaviour of the partial sample means, $\bar{z}_k$, of the process that are defined as

$$\bar{z}_k = \frac{(z_{k-1} + z_{k-2} + \cdots + z_o)}{k}. \qquad [11.5.1]$$

For a short memory time series, a plot of $\bar{z}_k$ against k should show great stochastic variability for the first values of k, but after k reaches a moderate value the graph should decay to an almost constant value and should show small stochastic variability. However, for the case of a long memory time series, the plot of $\bar{z}_k$ against k should display great stochastic variability for the first few values of k. For moderate values of k the graph should show a gentle trend that should oscillate around a constant value as k increases, and after k reaches a very large value, which depends on the degree of persistence, $\bar{z}_k$ should reach a constant value. Furthermore, because the present values of the time series are correlated with the past, the current values of $\bar{z}_k$ are highly correlated with the past and, therefore, a plot of $\bar{z}_k$ against k could show local trends. To detect persistence, the rate of decay towards a constant value of the local trends is of interest, as is also the presence of an overall gentle trend. However, the presence of local trends in the plot of $\bar{z}_k$ against k by itself does not indicate the presence of persistence. Within the fourteen data sets, the St. Lawrence Riverflows have an overall decreasing trend. This trend is gentle enough to assume that it could be due to the presence of persistence in the time series and not due to nonstationarity. The graph of $\bar{z}_k$ against k of the St. Lawrence Riverflows is displayed in Figure 11.5.1. For some of the other data sets, local trends in $\bar{z}_k$ seemed to be present even at the end of the series. Finally, for most of the data sets the behaviour of the partial means is consistent with what could be expected in time series having a short term memory, consisting of a rapid decay of the graph to a constant value.

All of the FARMA models considered for fitting to the series in Table 11.5.1 are subsets of the FARMA(2,d,1) model given by

$$(1 - \phi_1 B - \phi_2 B^2)\nabla^d(B)(z_t - \mu) = (1 - \theta_1 B)a_t, \qquad [11.5.2]$$

where ϕ_i is the ith AR parameter, θ_1 is the first MA parameter, and μ is not present in [11.5.2] for positive d. For the St. Lawrence Riverflows the additional constrained AR(3) model given by

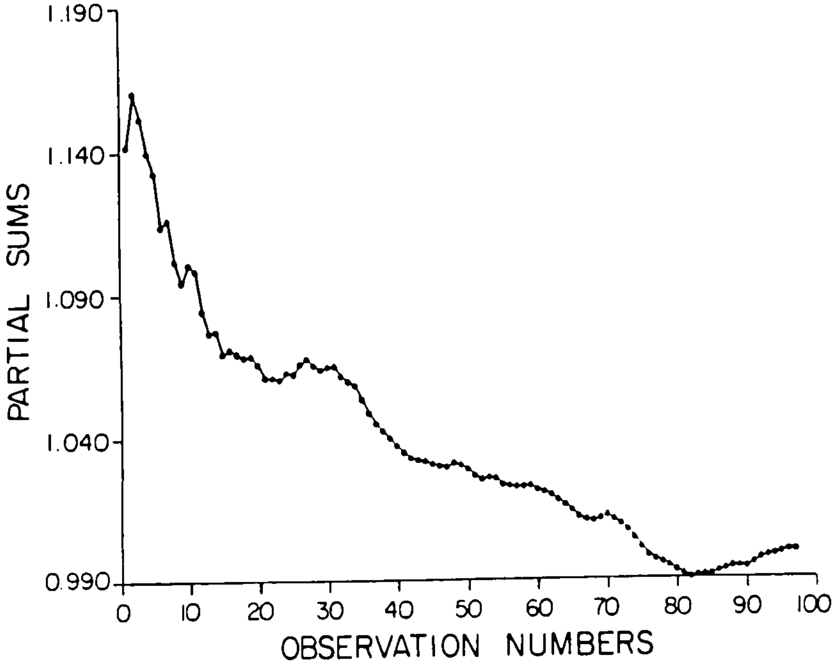


Figure 11.5.1. Partial sums of the St. Lawrence at Ogdensburg, New York from 1860-1957.

$$(1 - \phi_1 B - \phi_3 B^3)(z_t - \mu) = a_t, \qquad [11.5.3]$$

was considered, because this is the model used in Chapter 3 and Part III, within the class of ARMA models. The most appropriate FARMA model from [11.2.4] to fit each series was selected according to the minimum AIC (see Section 6.3), considering only those models that passed tests for whiteness of the fitted model residuals. The maximum likelihood estimates (MLE's) of the model parameters for each series along with the standard errors (SE's) given in brackets are displayed in Table 11.5.2. Those time series for which the estimates of d given in Table 11.5.2 are positive, portray persistent behaviour. Also, because the degree of persistence depends on the magnitude of d, those series having higher values of d possess greater degrees of persistence. For example, the model for the St. Lawrence River was estimated as a FARMA(0,d,0) with $d = 0.4999$. This indicates that the flows of the St. Lawrence are highly persistent, and, hence the far away past strongly influences the present. A consequence of this influence is the slow rate of convergence of the sample mean to the true value. For the case of the St. Lawrence River this rate of decay is of order $O(N^{-0.0001})$, where N is the number of observations. This order of convergence is also true for the forecasting function and the estimated ACF of the St. Lawrence Riverflows. An interesting feature of the St. Lawrence River is that it is associated with great masses of water which perhaps suggests a model having a reservoir term whose time step is larger than the time step used to measure the series. All the models that exhibit persistence in Table 11.5.2 are FARMA(0,d,0). The data sets for which it was appropriate to fit a FARMA(0,d,0) model are the Mckenzie, St. Lawrence and Thames annual riverflows plus the Philadelphia rainfall series. There were other data sets for which the AIC selected ARMA models but the differences between minimum AIC's for the ARMA models and the AIC's for FARMA(0,*d*,0) models were very small. Finally, note that some rivers do not show any sign of second order correlation structure, as the optimal model according to the AIC was simply the mean. These data sets are the Dal, Danube and Rhine Rivers. Keep in mind that most of these findings are consistent with the models suggested by the sample ACF, sample PACF and behaviour of the partial means.

Table 11.5.2. Parameter estimates and standard errors in brackets for FARMA models fitted to the hydrological time series.

Series	Parameter Estimates and SE's			
	ϕ_1	ϕ_2	d	θ_1
Saugeen	-	-	-	-
Dal	-	-	-	-
Danube	-	-	-	-
French	-0.234	-	-	-
	(0.12)			
Gota	0.59	-0.27	-	-
	(0.08)	(0.08)		
Mckenzie	-	-	0.27	-
			(0.10)	
Mississippi	0.29	-	-	-
	(0.10)			
Neumunas	-	-	-	-0.19
				(0.08)
Rhine	-	-	-	-
St. Lawrence	-	-	0.499	-
			(0.08)	
Thames	-	-	0.12	-
			(0.10)	
Fortaleza	0.24	-	-	-
	(0.08)			
Philadelphia	-	-	0.23	-
			(0.08)	
Temperature	0.12	0.2	-	-
	(0.06)	(0.06)		

The bootstrapping technique of Cover and Unny (1986) was used to increase the finite sample information about the estimates of the persistence parameter *d*. For some data sets, the AIC does not provide a clear cut separation between models with and without the persistence parameter *d*. Also, most of the time the best FARMA model with a persistent parameter was the FARMA(0,*d*,0) model. Because of these two remarks, it is interesting to obtain information on how reliable are the estimates of *d* and the estimates of its SE. For these reasons, the bootstrap technique proposed by Cover and Unny was used with the FARMA(0,*d*,0) model for all the data sets. Although this model is not appropriate for some data sets, the information about the behaviour of the estimates of *d* is valuable. The information given by the bootstrap was summarized by two methods. First, the sample mean and standard deviation of the estimates of d using the bootstrap technique for each data set were computed. Second, the distribution of the estimates of d for each data set was estimated using a nonparametric kernel estimate (Fryer, 1977). Using these two pieces of information, it is possible to decide if the data set exhibits any evidence of persistence. The means and standard deviations are given in Table 11.5.3, together with the estimates of d obtained using the approximate maximum likelihood method. Plots of the density of the estimates of d for the St. Lawrence riverflows and for rainfall at Philadelphia are given in Figures 11.5.2 and 11.5.3, respectively. From these tables and graphs, one can draw a number of conclusions:

(a) The large sample approximations are not necessarily valid for finite sample sizes. For example, the distribution of the estimate of the parameter d for the St. Lawrence seems not to have tails, and the density seems to be concentrated on the interval 0.4 to 0.54. Moreover, the density seems somewhat skewed.

(b) It appears that the asymptotic standard deviations are smaller than the bootstrap estimates for values of d not very close to 0.5, and this behaviour is reversed for values of d close to 0.5.

(c) Note also that the means of the estimates of the parameter d obtained by resampling and the standard deviations of these estimates do not necessarily represent the data.

If the threshold a is assumed known, the estimate of a by least squares can be easily found, and it can be shown using standard techniques that it has an asymptotic normal distribution with the inverse of the information matrix as the asymptotic variance. This information matrix is given by the expected value of the truncated variable $z_t^2\{|z_t| > a\}$. If the threshold a is unknown, the usual techniques cannot be used because of the nondifferentiability of the sum of squares function with respect to a.

11.6 CONCLUSIONS

As a direct result of research on long memory modelling motivated by the controversy surrounding the Hurst phenomenon defined in Section 10.3.1, Hosking (1981) originally proposed the generalization of ARIMA models so that long term persistence could be effectively modelled. In particular, Hosking (1981) and independently, Granger and Joyeux (1980), suggested the FARMA(p,d,q) model in [11.2.4] as a flexible approach for describing persistence. The FARMA model is especially appealing to researchers and practitioners in hydrology, economics and elsewhere, because it can model both long and short term behaviour within the confines of a single model. Bloomfield (1992), for example, employs FARMA models for investigating trends in annual global temperature data. The fractional differencing filter in [11.2.3] can account for long term behaviour or persistence while the ARMA component of the overall FARMA model in [11.2.4] takes care of the short memory aspects of the series being modelled. Because of these and other reasons, the FARMA(p,d,q) model of this chapter constitutes a more flexible approach to modelling persistence than the FGN model of Section 10.4.

Model construction techniques are available for fitting FARMA(p,d,q) models to data sets. However, as noted in Section 11.3 improved estimation techniques should be developed and further contributions to model identification could be made. An approximate maximum likelihood estimation algorithm is presented in Appendix A11.1. The applications of Section 11.5 demonstrate how FARMA(0,d,0) models can be fitted in practice to yearly time series. After obtaining MLE's for the model parameters and subjecting the fitted model to diagnostic checks

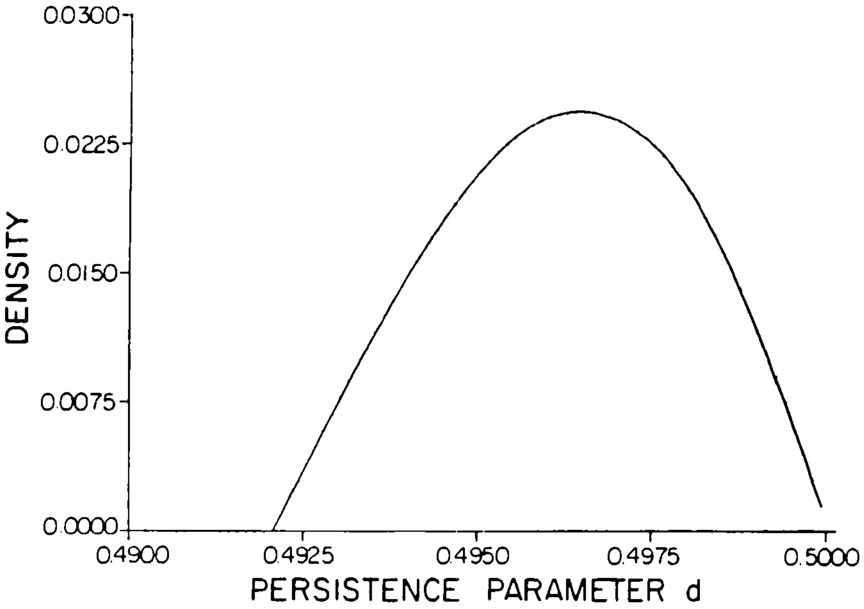


Figure 11.5.2. Probability density of the persistence parameter *d* obtained by bootstrapping for the St. Lawrence Riverflows.

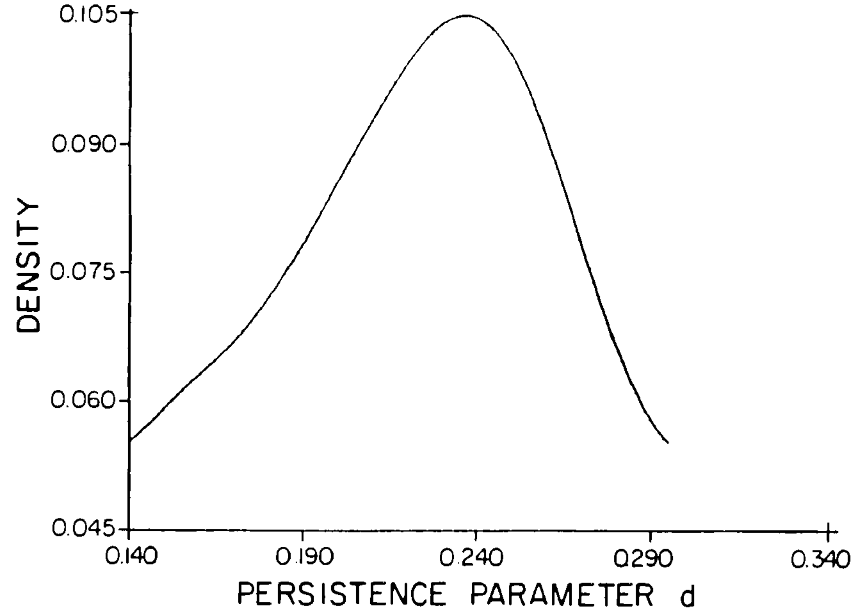


Figure 11.5.3. Probability density of the persistence parameter *d* obtained by bootstrapping for the Philadelphia rainfall.

Table 11.5.3. Estimation of the parameter d using bootstrapping.

Data Set Identification	Means of $\hat{d}$	St. Deviations	$\hat{d}$	St. Deviations
(1) Saugeen	0.110	0.212	0.108	0.100
(2) Dal	0.028	0.177	0.024	0.093
(3) Danube	0.069	0.157	0.059	0.072
(4) French	0.148	0.168	0.134	0.093
(5) Gota	0.365	0.245	0.388	0.634
(6) Mckenzie	0.234	0.142	0.274	0.105
(9) Neumunas	0.105	0.137	0.103	0.068
(10) Rain Phil.	0.210	0.110	0.229	0.078
(12) St. Lawrence	0.475	0.055	0.499	0.079
(13) Thames	0.139	0.149	0.120	0.093
(14) Temperature	0.153	0.079	0.151	0.050

Using the bootstrapping technique described in Section 11.3.3, the value of the persistence parameter d in the model $\nabla^d(B)z_t = a_t$ was estimated by the mean value of the estimates obtained using the bootstrap, and is given in the second column. The third column gives the standard deviation of the estimate obtained by bootstrapping. The fourth column lists $\hat{d}$ which is the estimate of d obtained by the appropriate maximum likelihood method described in the Section 11.3.3 and the last column gives the asymptotic standard deviation of $\hat{d}$.

(Section 11.3.4), the calibrated model can be used for simulation and forecasting. Techniques for simulating and forecasting with a FARMA(p,d,q) model are presented in Section 11.4.

APPENDIX A11.1

ESTIMATION ALGORITHM FOR FARMA MODELS

This appendix presents an algorithm for obtaining approximate MLE's for the parameters of a FARMA(p,d,q) model. This estimation algorithm was originally presented by Jimenez et al. (1990) and constitutes an extension of the estimation algorithm of Li and McLeod (1986). In the algorithm, it is assumed that the estimated mean of the series has been subtracted from each observation in order to produce a series having a zero mean. The mean can be estimated using [11.3.2] or some other appropriate technique.

To compute the unconditional sum of squares of the residuals obtained assuming that the model can be represented by a long autoregressive approximation of

$$\phi(B)\nabla_M^d(B)z_t = \theta(B)a_t, \qquad [A11.1.1]$$

a backforecasting algorithm similar to the one used by McLeod and Sales (1983) to compute the unconditional residual sum of squares for seasonal ARMA models, can be used. The unconditional sum of squares of the residuals is given by

$$S = \sum_{t=-\infty}^{N} [a_t]^2, \qquad [A11.1.2]$$

where $[\cdot]$ denotes expectation with respect to the observations is approximated by

$$S = \sum_{t=1-Q}^{N} [a_t]^2, \qquad [A11.1.3]$$

where Q is a fairly large truncation point. The conditional form of [A11.1.1] is given by

$$\phi(B)\nabla_M^d(B)[z_t] = \theta(B)[a_t], \qquad [A11.1.4]$$

where $[a_t] = 0$, $t > N$. This can be expressed by a two stage model

$$\nabla_M^d(B)[z_t] = [c_t], \qquad [A11.1.5]$$

and

$$\phi(B)[c_t] = \theta(B)[a_t] \qquad [A11.1.6]$$

The Box-Jenkins backforecasting approach needs also the forward form of [A11.1.1] such that

$$\phi(F)\nabla_M^d(F)z_t = \theta(F)e_t, \qquad [A11.1.7]$$

where F is the forward time shift operator such that $Fz_t = z_{t+1}$ and e_t is a sequence of normal independent random variables with mean 0. The method uses the conditional form of [A11.1.7] given by

$$\nabla_M^d(F)[z_t] = [b_t], \qquad [A11.1.8]$$

and

$$\phi(F)[b_t] = \theta(F)[e_t] \qquad [A11.1.9]$$

where $[e_t] = 0$, $t < 1$.

In summary, the unconditional sum of squares can be obtained iteratively through the following steps.

Step 0. Select Q and M.

Step 1. Compute the autoregressive coefficients of ∇_M^d.

Step 2. Compute $[b_t]$, using [A11.1.9] for $t = N + Q, \ldots, 1$. Initially set $[b_t] = 0$.

Step 3. Backforecast the $[b_t]$ series using [A11.1.9]. This can be accomplished using the SARMAS algorithm of McLeod and Sales (1983).

Step 4. Backforecast the $[z_t]$ series using [A11.1.8].

Step 5. Compute the $[c_t]$ for $t = 1 - Q, \ldots, N$ series using [A11.1.5].

Step 6. Compute the $[a_t]$ for $t = 1 - Q, \ldots, N$ series using [A11.1.6].

Step 7. Compute S using [A11.1.2].

Steps 1 to 7 can be repeated until a previously specified tolerance limit is achieved. The parameters are obtained by minimizing S as given in [A11.1.2]. The minimization algorithm given by Powell (1964) can be employed to minimize S.

PROBLEMS

11.1 By referring to appropriate literature cited in Chapters 10 and 11, make a list of the range of related definitions for long memory, or persistence. Compare the similarities and differences among these definitions. Which definition is most clear to you?

11.2 The definition for a FARMA(p,d,q) model is presented in [11.2.4]. Employ [11.2.3] to write the expanded forms of the following FARMA models:

a) FARMA(1,0.4,1),

b) FARMA(0,-0.3,2),

c) FARMA(1,0.8,1).

11.3 Long memory models have been applied to time series in a variety of different fields. Find three different applications of long memory models by referring to publications in fields of your choice. For each application, write down the complete reference and provide a brief summary. Do not use applications from references given in Chapters 10 and 11.

11.4 By referring to the paper by Hosking (1981), write down the formula for the theoretical ACF of a fractional differencing model and comment upon the general properties of the ACF.

11.5 Outline the purposes of bootstrapping and how it is implemented in practice. Describe in some detail how the bootstrapping technique of Cover and Unny (1986) can be employed when estimating parameters in ARMA, ARIMA and FARMA models.

11.6 Select an annual time series which you think may possess long term memory. Explain reasons for suspecting persistence based upon your physical understanding of the problem. Following the approaches suggested in Sections 11.3 and 11.5, use statistical identification methods to justify your suspicions. Fit a fractional differencing or FARMA(0,d,0) as well as the most appropriate ARMA model to the series. Comment upon your findings.

REFERENCES

BOOTSTRAPPING

Cover, K. A. and Unny, T. E. (1986). Application of computer intensive statistics to parameter uncertainty in streamflow synthesis. *Water Resources Bulletin*, 22(3):495-507.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1-26.

Fryer, M. J. (1977). A review of some nonparametric methods of density estimation. *Journal of the Institute of Mathematics and Applications*, 20:335-354.

FARMA MODELLING

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change* 21:1-16.

Boes, D. C., Davis, R. A. and Gupta, S. N. (1989). Parameter estimation in low order fractionally differenced ARMA processes. *Stochastic Hydrology and Hydraulics*, 3:97-110.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer-Verlag, New York.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, 68:165-176.

Hosking, J. R. M. (1984). Modeling persistence of hydrological time series using fractional differencing. *Water Resources Research*, 20(12):1898-1908.

Hosking, J. R. M. (1985). Fractional differencing modelling in hydrology. *Water Resources Bulletin*, 21(4):677-682.

Hui, Y. V. and Li, W. K. (1988). On fractional differenced periodic processes. Technical report, Chinese University of Hong Kong.

Jimenez, C., Hipel, K. W. and McLeod, A. I. (1990). D´evelopements recents dans la modelisations de la persistance à long terme. *Revue des Sciences de l'Eau*, 3(1):55-81.

Li, W. K. and McLeod, A. I. (1986). Fractional time series modelling. *Biometrika*, 73(1):217-221.

LONG MEMORY

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

Cox, D. R. (1984). Long range dependence: A review. In *Statistics: An Appraisal*, David, H. A. and David, H. T., editors, Proceedings of the 50th Anniversary, Iowa State Statistical Laboratory, Iowa State University Press, Ames, Iowa, pp. 55-74.

Davison, A. C. and Cox, D. R. (1989). Some simple properties of sums of random variables having long-range dependence. *Proceedings of the Royal Society of London*, A424:255-262.

Eberlein, E. and Taqqu, M. S., editors (1986). *Dependence in Probability and Statistics: A Survey of Recent Results*. Birkhauser, Boston.

Granger, C. W. J. (1980). Long-memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14:227-238.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1(1):15-29.

Rosenblatt, M. (1961). Independence and dependence. In *Proceedings of the 4th Berkley Symposium on Mathematical Statistics and Probability.* 431-443.

Rosenblatt, M. (1979). Some limit theorems for partial sums of quadratic forms in stationary Gaussian variables. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 49:125-132.

Rosenblatt, M. (1981). Limit theorems for Fourier transforms of functional Gaussian sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 55:123-132.

Taqqu, M. (1975). Weak convergence to fractional Brownian motion and the Rosenblatt process. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 31:287-302.

TIME SERIES ANALYSIS

Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society,* Series B, 8:27-41.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Cryer, J. D. and Ledolter, J. (1981). Small-sample properties of the maximum likelihood estimator in the first-order moving average model. *Biometrika,* 68:691-694.

Durbin, J. (1960). The fitting of time series models. *Revue de L'Institut International de Statistique,* 28(3):233-244.

Healy, M. J. R. (1968). Algorithm AS6, triangular decomposition of a symmetric matrix. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 17:195-197.

McLeod, A. I. and Hipel, K. W. (1978a). Comment on modelling monthly hydrologic persistence by G. K. Young and R. U. Jettmar. *Water Resources Research,* 14(4):699-702.

McLeod, A. I. and Hipel, K. W. (1978b). Simulation procedures for Box-Jenkins models. *Water Resources Research,* 14(5):969-975.

McLeod, A. I. and Sales, P. R. H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 32:211-223.

Parzen, E. (1982). ARARMA models for time series analysis and forecasting. *Journal of Forecasting,* 1:67-82.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal,* 7:155-162.

Trench, W. F. (1964). An algorithm for the inversion of finite Toeplitz matrices. *SIAM Journal of Applied Mathematics,* 12:515-521.

PART VI

SEASONAL MODELS

At a given location on the earth's surface, **hydrologic phenomena**, as well as other natural occurrences tend to vary from one season to another. In many cases, the change in behaviour of a specified phenomenon can be quite spectacular. For example, in Northern Canada the winters are extremely cold whereas the summer temperatures are quite moderate. In addition, due to the melting of snow during the early springtime, the spring flows of northern rivers tend to be quite high while the winter flows are much less. The foregoing and other kinds of hydrologic changes across seasons, are largely dependent upon the rotation of the earth about the sun along with the accompanying changes in tilt of the earth's axis. Local conditions such as mountain ranges, large bodies of water and continental land masses, also influence the hydrologic characteristics of a region.

Seasonal hydrological data are available as time series for which the time intervals between adjacent observations are less than one year. For instance, natural time series are commonly available as average daily, weekly, and monthly sequences. As an illustration of an average monthly time series consider Figure VI.1 which shows the last ten years of the **average monthly flows** of the Saugeen River at Walkerton, Ontario, which are available from January 1915 until December 1976 (Environment Canada, 1977). The sinusoidal shape of this graph is caused by the seasonal changes of the flows throughout the year. As can be seen, the riverflows are high during the spring months and much lower in the winter time. Because the flows vary across the seasons or periods in a cyclic manner, the time series in Figure VI.1 is often referred to as a periodic series.

Except for some random variation, the flows within a given season in Figure VI.1 tend to be stationary across the years. For example, the individual April flows of the Saugeen River vary about the April mean across all of the years from 1915 to 1976. There appears to be no upward or downward trend across the years in the April flows. However, not all periodic series are stationary within each season. Often series which are dependent upon social and economic behaviour possess statistical characteristics which vary across the years within each season. Consider, for example, the graph in Figure VI.2 of the **average monthly water consumption** in millions of litres per day from 1966 to 1988 for the city of London, Ontario, Canada. These observations are available from the Public Utilities Commission (1989) of London. Like other prosperous cities in North America, the city of London has grown dramatically in size since World War II to a 1991 population of about 300,000 people. This increasing population growth, in conjunction with other socio-economic factors have caused the water demand to increase greatly with time, as portrayed by the obvious upward trend in Figure VI.2. Moreover, the seasonality is clearly visible as a sinusoidal pattern wrapped around the trend.

The **greenhouse effect** is referred to in Section 1.2.1. One of the principal greenhouse gases is carbon dioxide (CO_2) which could in conjunction with other gases lead to serious global warning. Figure VI.3 displays a graph of the average monthly concentrations of atmospheric CO_2 measured at the Mauna Loa Observatory located at 3397 meters above sea level on the shoulder of Mauna Loa on the Island of Hawaii. Because the Mauna Loa Observatory is remote from industrial and other man-made sources of CO_2, the CO_2 observations collected there are believed to be a reliable indicator of the regional trend in the concentration of atmospheric CO_2 in the middle layers of the troposphere. Moreover, the Mauna Loa CO_2 program constitutes the first continuous station anywhere to produce reliable CO_2 data which are listed by both Keeling et al. (1982) and Bacastow and Keeling (1981). As depicted in Figure VI.3 there is a distinct upward trend in the CO_2 data which is probably caused by man-induced interventions such as the large-scale burning of fossil fuels and the massive destruction of forests throughout the world. Consequently, this series provides a vivid illustration of how a natural phenomena can be significantly altered by civilization. One can also see in Figure VI.3 the periodic nature of the CO_2 data which fluctuate in a sinusoidal manner about the trend.

Within section VI, the following three families of seasonal models are presented for modelling seasonal data:

1. **seasonal autoregressive integrated moving average (SARIMA) models** (Chapter 12);
2. **deseasonalized models** (Chapter 13);
3. **periodic models** (Chapter 14).

The deseasonalized and periodic models are used for describing data such as the average monthly flows shown in Figure VI.1, which possess stationarity within each season. The SARIMA family of models can be fitted to data such as those shown in Figures VI.2 and VI.3 where the level and perhaps other statistics change within each season across the years.

The **designs** of all three classes of seasonal models constitute direct extensions of nonseasonal models to the seasonal or periodic cases. More specifically, the SARIMA family of models is a seasonal version of the ARIMA class of models described in Chapters 3 and 4 of Part II. For a given series such as the one in Figure VI.2 or VI.3, nonseasonal and seasonal nonstationarities are removed using nonseasonal and seasonal differencing operators, respectively, before fitting a stationary seasonal ARMA model to the series. To fit a deseasonalized model to a periodic series like the one in Figure VI.1, each observation in the series is first deseasonalized by subtracting out the seasonal mean and then dividing this by the seasonal standard deviation. A nonseasonal ARMA model from Chapter 3 can then be fitted to the resulting nonseasonal series. Finally, a periodic model is formed by fitting a separate PAR or AR model to each season of the year to form what are called **periodic autoregressive (PAR) models** or **periodic autoregressive-moving average (PARMA) models**, respectively. In this way, the varying correlation structure across the seasons in a series such as in the one shown in Figure VI.1 can be directly modelled.

In the next three chapters, each of the three types of seasonal models is defined and **model construction** procedures are presented. Hydrologic and other kinds of applications are employed for demonstrating how the models can be conveniently applied in practice to seasonal data. Because the deseasonalized and periodic models possess quite a few model parameters, procedures are given for reducing the number of model parameters. **Forecasting experiments** in Chapter 15, demonstrate that a certain type of PAR model forecasts seasonal hydrologic time series better than other kinds of competing seasonal models. Consequently, when sufficient data are available, periodic models are the best type of models to use for describing hydrologic and other kinds of natural time series.

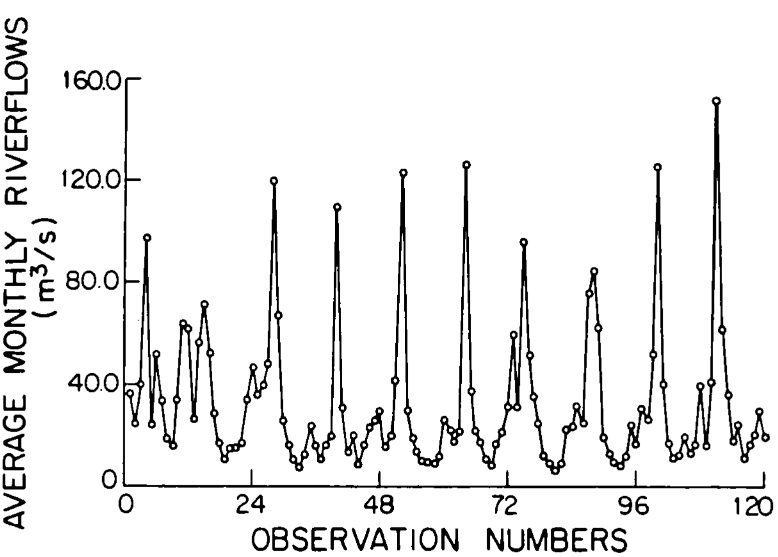


Figure VI.1. Average monthly flows (m^3/s) of the Saugeen River at Walkerton, Ontario, Canada, from January, 1967, until December, 1976.

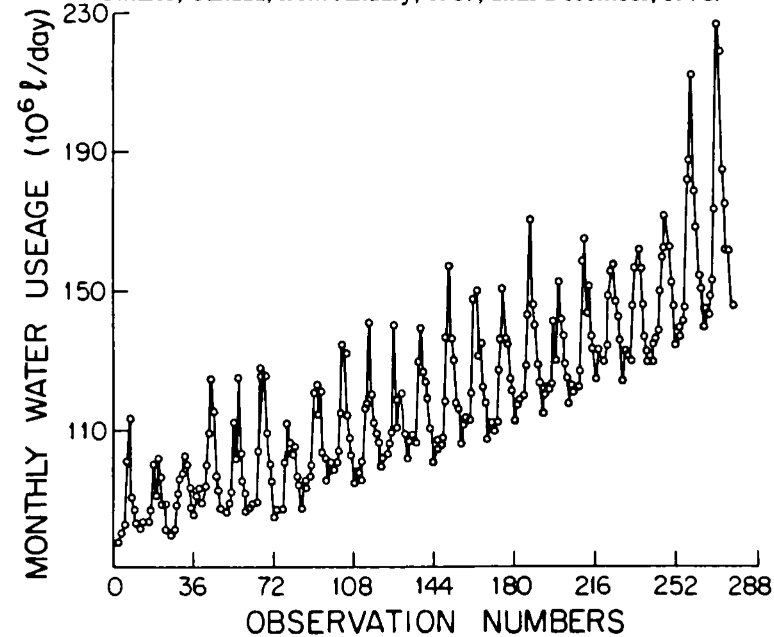


Figure VI.2. Average monthly water useage (million litres per day) for the city of London, Ontario, Canada, from January, 1966, to December, 1988.

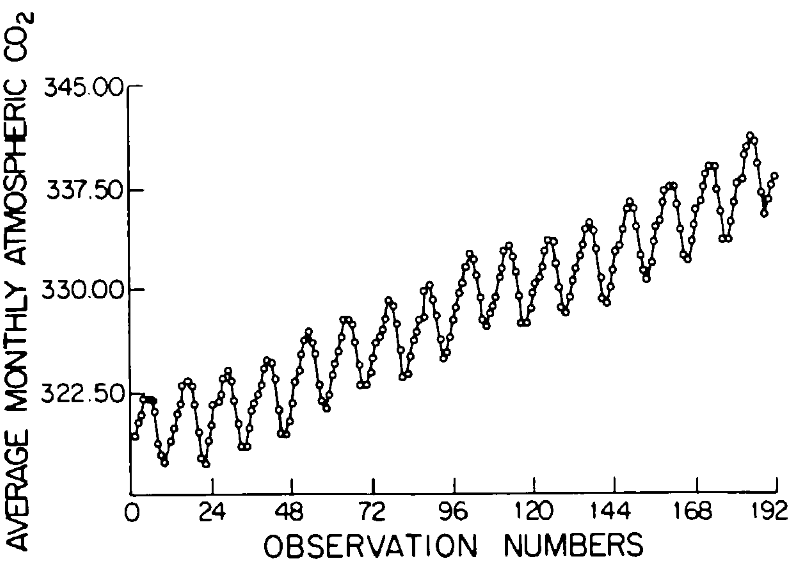


Figure VI.3. Average monthly concentrations of atmospheric CO_2 (molefraction in ppm) measured at Mauna Loa Observatory in Hawaii from January, 1965, to December, 1980.

Although all three seasonal families of models can be used for forecasting hydrologic series, only the deseasonalized and periodic models are properly designed for simulating the seasonally stationary type of data such as the time series shown in Figure VI.1. Additionally, only the SARIMA model contains the appropriate model parameters for describing the nonstationary seasonal data of Figures VI.2 and VI.3. If, for example, the residuals of a SARIMA model fitted to a seasonal series possess correlation which varies from season to season, a PAR model could be fitted to the residuals to capture this behaviour.

CHAPTER 12

SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

12.1 INTRODUCTION

Seasonal autoregressive integrated moving average (SARIMA) models are useful for modelling seasonal time series in which the mean and other statistics for a given season are not stationary across the years. The graphs of the average monthly water consumption and atmospheric CO_2 series displayed in Figures VI.2 and VI.3, respectively, depict revealing illustrations of this type of behaviour. Notice that both data sets possess increasing trends which in turn means that the level of each monthly observation is growing in magnitude within the same month over the years. Moreover, the sinusoidal curves that follow the upward trends confirm that the data are seasonal. Figures VI.2 and VI.3 are representative of the general types of nonstationary statistical characteristics that are often present in many kinds of socio-economic time series and natural time series that are significantly affected by man-induced changes, respectively. Other examples of time series which would behave in a similar fashion to the one in Figures VI.2 and VI.3 include average monthly irrigation water consumption, average weekly electricity demand and total quarterly income for recreational facilities located near lakes and rivers.

In the next section, the *mathematical design* of the SARIMA model is presented and associated theoretical properties are described. An inspection of this design indicates why the SARIMA model is ideally suited for modelling a seasonal nonstationary time series like those shown in Figures VI.2 and VI.3 using relatively few model parameters. However, because the mathematical definition does not contain model parameters which explicitly account for separate means and variances in each season, the SARIMA model is not suitably designed for describing series having stationarity of second order moments within each season across the years. For instance, because the average monthly riverflow series of the Saugeen River at Walkerton, Ontario, Canada, plotted in Figure VI.1, appears to have a seasonal mean and variance which are more or less stationary across all the years for each season, a SARIMA is not the best model to fit to this series. Rather, the *deseasonalized and periodic models* of Chapters 13 and 14, respectively, can be employed for modelling this series. Nonetheless, when one is confronted with modelling a seasonal series similar to the one in Figure VI.1, the model construction techniques of Section 12.3 can be utilized for conveniently fitting an appropriate SARIMA model to the series. The *applications* contained in Section 12.4 clearly explain how SARIMA models are fitted in practice by following the identification, estimation and diagnostic check stages of *model construction*. Subsequent to fitting the most appropriate SARIMA model to a series, the calibrated model can be used for purposes such as *forecasting and simulation*, as explained in Section 12.5.

12.2 MODEL DESIGN

12.2.1 Definition

The SARIMA model defined in this section constitutes a straightforward extension of the nonseasonal ARMA and ARIMA models presented in Chapters 3 and 4, respectively. In their book, Box and Jenkins (1976, Ch. 9) define this model and justify why it is useful for describing certain kinds of seasonal series.

Let $z_1, z_2, \ldots, z_\eta$, represent a sequence of seasonal observations. If, for example, there were n years of data for which each year contains s seasons, this would mean that η is equal to ns. For the case of 15 years of monthly data, there would be a total of $15 \times 12 = 180$ observations. If the seasonal time series were not normally distributed and/or the variance of the series changes over time (i.e., the series is heteroscedastic), one could alleviate this problem by invoking a *Box-Cox transformation* (Box and Cox, 1964) defined in [3.4.30] as

$$z_t^{(\lambda)} = \begin{cases} \lambda^{-1}[(z_t + c)^\lambda - 1], & \lambda \neq 0 \\ \\ \ln(z_t + c), & \lambda = 0 \end{cases} \qquad [12.2.1]$$

The parameter λ is the Box-Cox power transformation and c is a positive number which is chosen to be just large enough to cause all the entries in the time series to be positive. If nonnormality and heteroscedasticity in the given series were not detected prior to fitting a SARIMA model to the data, these problems would show up in the residuals of the fitted model. At that time, an appropriate Box-Cox transformation could be selected and the parameters of the SARIMA model could then be estimated for the transformed series.

Figures VI.2 and VI.3 graphically depicts how the magnitudes of observations can change across the seasons in a cyclic manner and also from year to year within a given season. To eliminate nonstationarity within each season, one can employ the *seasonal differencing operator* defined by

$$\nabla_s z_t^{(\lambda)} = (1 - B^s) z_t^{(\lambda)} = z_t^{(\lambda)} - z_{t-s}^{(\lambda)} \quad \text{for } t = s+1, s+2, \cdots, \eta \qquad [12.2.2]$$

where s is the number of seasons per year and B^s is the backward shift operator defined by $B^s z_t^{(\lambda)} = z_{t-s}^{(\lambda)}$. When dealing with monthly data, notice that the relationship in [12.2.2] only connects observations within the same season. Hence, when using seasonal differencing with monthly data, an observation in March is only subtracted from the observation in March of the previous year. If the $z_t^{(\lambda)}$ series is of length $\eta = sn$, the number of observations in the differenced series is $\eta - s$. The differencing operator in [12.2.2] is applied just enough times to remove the seasonal nonstationarity. If it were necessary to apply the seasonal differencing operator in [12.2.2] D times to produce a series of length $\eta - sD$, the resulting series would be given by

$$\nabla_s^D z_t^{(\lambda)} = (1 - B^s)^D z_t^{(\lambda)} \qquad [12.2.3]$$

For purposes of explanation, consider once again a time series consisting of monthly observations. To model correlation among, say, March observations in the differenced series, one may wish to introduce appropriate model parameters. More specifically, to accomplish this task of linking March observations together one can use a model of the form

$$\Phi(B^s)\nabla_s^D z_t^{(\lambda)} = \Theta(B^s)\alpha_t \qquad [12.2.4]$$

where $\Phi(B^s)$ and $\Theta(B^s)$ are the seasonal autoregressive (AR) and seasonal moving average (MA) operators, respectively, and α_t is a residual series which may contain nonseasonal correlation. Both the seasonal AR and MA operators are defined in order to describe relationships within the same season. In particular, the *seasonal AR operator* is defined as

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

where Φ_i is the ith AR parameter and P is the order of the AR operator. Because the power of each differencing operator is always an integer multiple of s, only the observations within each season are related to one another when using this operator. Hence, for the case of March observations in a monthly series, only the March observations are connected together using $\Phi(B^s)$. To describe the relationship of the residuals, α_t, within a given season, the *seasonal MA operator* is defined using

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$$

where Θ_i is the ith MA parameter and Q is the order of the MA operator. Since the exponents of B in $\Theta(B^s)$ are always integer multiples of s, the residuals in the same season are linked with another when using $\Theta(B^s)$.

Theoretically one could define a separate model as in [12.2.4] for each season of the year. However, to keep the model as parsimonious as possible, one can assume that [12.2.4] can be used for all of the seasons. Therefore, one is making the assumption that the correlation within all of the seasons is the same. For the case of monthly data this means that the relationship among all of the March observations is exactly the same as each of the other months.

The error components or residuals, α_t, may contain nonseasonal nonstationarity which can be removed by using the *nonseasonal differencing operator* defined in [4.3.3] as

$$\nabla^d \alpha_t = (1 - B)^d \alpha_t \qquad [12.2.5]$$

where d is the order of the nonseasonal differencing operator which is selected just large enough to remove all of the nonseasonal nonstationarity. The sequence produced using [12.2.5] is theoretically a stationary nonseasonal series. The nonseasonal correlation can then be captured by writing the ARMA model in [3.4.4] or [4.3.4] as

$$\phi(B)\nabla^d \alpha_t = \theta(B) a_t \qquad [12.2.6]$$

where $\phi(B)$ is the *nonseasonal AR operator* of order p defined as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

and $\theta(B)$ is the *nonseasonal MA operator* of order q written as

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

The a_t's are the innovations which are identically independently distributed (IID) with a mean of zero and variance of σ_a^2. Hence $a_t \sim IID(0,\sigma_a^2)$. In order to obtain maximum likelihood estimates for the model parameters of a SARIMA model, in the next section the restriction of normality is also placed upon the a_t's so that they are assumed to be distributed as $NID(0,\sigma_a^2)$.

Because of the form of [12.2.6], the correlation among seasons is the same no matter what season one is dealing with. Hence, when entertaining monthly data, the correlation between, say, the March and February observations is defined to be the same as that between any other adjacent months such as October and September. For the periodic models of Chapter 14, this restriction is dropped by allowing for a separate correlation structure for each season of the year.

To define the overall seasonal model, one simply combines equations [12.2.6] and [12.2.4]. This can be accomplished by solving for α_t in [12.2.6] and substituting this result into [12.2.4] to obtain the *SARIMA (seasonal autoregressive integrated moving average) model*

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D z_t^{(\lambda)} = \theta(B)\Theta(B^s)a_t \qquad [12.2.7]$$

Because the operators in [12.2.7] are multiplied together rather than summed, this model is often called a multiplicative SARIMA model.

When fitting the SARIMA model to a given time series of length η, one first transforms the data, if necessary, using the Box-Cox transformation in [12.2.1]. Following this, the data can be differenced both seasonally and nonseasonally. It does not matter which differencing operation is carried out first. One then obtains the stationary series given by

$$w_t = \nabla^d\nabla_s^D z_t^{(\lambda)} \qquad [12.2.8]$$

where the length of the w_t series is $\eta' = \eta - d - sD$. The seasonal and nonseasonal correlation in the w_t sequence is modelled by using the seasonal and nonseasonal AR and MA operators, respectively. Hence, w_t is modelled by employing

$$\phi(B)\Phi(B^s)w_t = \theta(B)\Theta(B^s)a_t \qquad [12.2.9]$$

In some applications, w_t may be a stationary seasonal series which is not obtained by differencing the original series. The model in [12.2.9] is called a *seasonal ARMA or SARMA model* of the w_t series.

12.2.2 Notation

For a given application, one may first wish to indicate the key parameters included in a SARIMA model, without writing down all of the parameter estimates either in a table or else using the difference equation in [12.2.7]. An economical notation for summarizing the structure of the SARIMA model in [12.2.7] is $(p,d,q)\times(P,D,Q)_s$. The first set of brackets contains the orders of the nonseasonal operators while the orders of the seasonal operators are listed inside the second set of brackets. More specifically, p, d and q stand for the orders of the nonseasonal AR, differencing and MA operators, respectively. In the second set of brackets, P, D and Q give the orders of the seasonal AR, differencing and MA operators, respectively. The subscript s appearing to the right of the second set of brackets points out the number of seasons per year.

For the case of monthly data for which $s = 12$, a specific example of a SARIMA model is $(2,1,1)\times(1,1,2)_{12}$. Suppose that the original series were transformed using natural logarithms. By utilizing [12.2.7], this model is written using a finite difference equation as

$$(1-\phi_1B-\phi_2B)(1-\Phi_1B^{12})(1-B)(1-B^{12})\ln(z_t)=(1-\theta_1B)(1-\Theta_1B^{12}-\Theta_2B^{24})a_t$$

If the data are stationary, nonseasonal or seasonal differencing is not required. A stationary model is indicated as $(p,0,q)\times(P,0,Q)_s$. Because this model is stationary, sometimes it is referred to as a SARMA (i.e., seasonal autoregressive-moving average) $(p,q)\times(P,Q)_s$ model.

The summary notation for a pure MA model is $(0,d,q)\times(0,D,Q)_s$. When a model contains no MA parameters, the SARIMA model is written as $(p,d,0)\times(P,D,0)_s$.

When a model is purely nonseasonal, the notation of Part II should be used. Hence, a stationary nonseasonal ARMA model is simply indicated by ARMA(p,q) instead of SARMA(p,q)×$(0,0)_1$. Likewise, a nonstationary nonseasonal ARIMA model is denoted as ARIMA(p,d,q) rather than the more cumbersome notation given by SARIMA(p,d,q)×$(0,0,0)_1$.

12.2.3 Stationarity and Invertibility

For a nonseasonal model, the conditions of stationarity and invertibility are discussed in detail in Sections 3.2.2, 3.3.2, and 3.4.2. Recall that for an ARMA model to be stationary the roots of the characteristics equation $\phi(B)=0$ must lie outside the unit circle. Likewise, for invertibility the roots of $\theta(B)=0$ must fall outside the unit circle.

In addition to the aforesaid conditions, the properties of the seasonal AR and MA operators must be specified for the SARIMA model if it is to be fitted to the stationary w_t series in [12.2.9]. For *seasonal stationarity* the roots of the characteristic equation $\Phi(B^s)=0$ must lie outside the unit circle. Similarly, for *seasonal invertibility*, the roots of the characteristic equation $\Theta(B^s)=0$ must fall outside the unit circle.

12.2.4 Unfactored and Nonmultiplicative Models

The SARIMA(p,d,q)×$(P,D,Q)_s$ model in [12.2.7] is referred to as a *multiplicative model*. This is because the nonseasonal and seasonal AR operators are multiplied together on the left hand side while the two MA operators are multiplied together on the right hand side. In addition, the nonseasonal and seasonal differencing operators are multiplied together with the AR operators in [12.2.7]. Following similar arguments, the SARMA(p,q)×$(P,Q)_s$ model which can be fitted to the stationary w_t series in [12.2.9] is also multiplicative.

Rather than write the SARIMA or SARMA model in multiplicative form, it is sometimes useful to use other formats. One approach is to write the models in unfactored form. This is accomplished by multiplying the $\phi(B)$ and $\Phi(B^s)$ operators together to create a single AR operator which can be labelled as $\bar{\phi}(B)$. Likewise, one can multiply $\theta(B)$ and $\Theta(B^s)$ together to form the single MA operator $\bar{\theta}(B)$. The SARIMA model in [12.2.7] can then be written in *unfactored form* as

$$\bar{\phi}(B)\nabla^d\nabla_s^Dz_t^{(\lambda)}=\bar{\theta}(B)a_t \qquad [12.2.10]$$

while the unfactored SARMA model for the w_t series is

$$\tilde{\phi}(B)w_t^{(\lambda)} = \tilde{\theta}(B)a_t \qquad [12.2.11]$$

As mentioned in Section 12.5, the unfactored form of the model in [12.2.10] or [12.2.11] is used for simulating data when differencing operators are present.

As an example of how an unfactored model is determined, consider a SARMA(2,1)×(1,2)$_s$ model which is written in multiplicative form as

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^s)w_t = (1 - \theta_1 B)(1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$$

The unfactored version of this model is

$$(1 - \phi_1 B - \phi_2 B^2 - \Phi_1 B^s + \phi_1 \Phi_1 B^{s+1} + \phi_2 \Phi_1 B^{s+2})w_t$$
$$= (1 - \theta_1 B - \Theta_1 B^s + \theta_1 \Theta_1 B^{s+1} - \Theta_2 B^{2s} + \theta_1 \Theta_2 B^{2s+1})a_t$$

In the unfactored model in [12.2.10], the nonseasonal and seasonal differencing operators are not included in the unfactored AR operator $\tilde{\phi}(B)$. To obtain a SARIMA model that does not have separate differencing factors, one can write the model in *generalized form* as

$$\phi'(B)z_t^{(\lambda)} = \theta'(B)a_t \qquad [12.2.12]$$

where $\phi'(B) = \phi(B)\Phi(B^s)\nabla^d\nabla_s^D$ is the *generalized or unfactored AR operator* for which ϕ'_i is the ith generalized AR parameter, and $\theta'(B) = \theta(B)\theta(B^s) = \tilde{\theta}(B)$ is the *generalized or unfactored MA operator* for which θ'_i is the ith generalized MA parameter. The unfactored format in [12.2.12] is useful for calculating minimum mean squared error forecasts, as is pointed out in Section 12.5.

Because the unfactored models are equivalent to their multiplicative counterparts, unfactored models are, of course, multiplicative. For some applications one may wish to generalize the multiplicative models given in equations [12.2.7], [12.2.9], [12.2.10] and [12.2.11], by allowing nonmultiplicative AR and MA operators, which are denoted by $\phi^*(B)$ and $\theta^*(B)$, respectively. If one wishes to remove nonseasonal and seasonal stationarity using differencing, the *nonmultiplicative model* is given as

$$\phi^*(B)\nabla^d\nabla_s^D z_t^{(\lambda)} = \theta^*(B)a_t \qquad [12.2.13]$$

When the given series is already stationary due to differencing or some other type of operation, the nonmultiplicative model is written as

$$\phi^*(B)w_t = \theta^*(B)a_t \qquad [12.2.14]$$

The model in [12.2.14] is, in fact, the same as the nonseasonal ARMA model given in [3.4.4]. However, in practice one would not use all of the parameters in the ARMA model and, therefore, use the constrained type of model discussed in Section 3.4.4. An example of a *constrained model* is

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_s B^s - \phi_{s+1} B^{s+1})w_t = (1 - \theta_1 B - \theta_s B^s - \theta_{s+1} B^{s+1})a_t$$

Notice that this model cannot be factored to form a multiplicative model, as is the case for the earlier example for the unfactored model.

In addition to entertaining the nonmultiplicative model, certain problems may dictate that a more complex multiplicative seasonal model should be used. For example, suppose that a water demand series contains two distinct seasonal components. One component may be due to industrial use which follows a different seasonal pattern than the second component, which is the residential demand. For each seasonal component, separate AR and MA operators could be defined and all of these operators would be combined together in an overall multiplicative seasonal model.

12.2.5 Autocorrelation Function

The derivations of the theoretical ACF for nonseasonal AR, MA, and ARMA processes are presented in Sections 3.2.2, 3.3.2 and 3.4.2, respectively. One could follow a similar approach to that used for the nonseasonal case to develop the formula for the theoretical ACF of the stationary seasonal process, w_t, given in [12.2.9]. However, a simpler approach is to write the SARMA model for w_t in [12.2.9] in the same format as its nonseasonal counterpart by using the unfactored form in [12.2.11]. Subsequent to this, the procedure developed for the nonseasonal case can be used to obtain the theoretical ACF for the seasonal model. To use the theoretical results for the ACF developed for the nonseasonal ARMA model in Section 3.4.2, simply replace the nonseasonal AR and MA operators by their combined counterparts for the seasonal model. The algorithm of McLeod (1975) presented in Appendix A3.2 can then be used to determine the theoretical ACF for the SARMA process. In a similar fashion, one can also calculate the theoretical PACF for the seasonal model by utilizing appropriate results developed for the nonseasonal case in Chapter 3.

12.2.6 Three Formulations of the Seasonal Processes

Introduction

The three formulations for the ARMA and ARIMA processes are given in Sections 3.4.3 and 4.3.4, respectively. As explained in these sections, the ARMA and ARIMA processes can be written using the original forms for their finite difference equations, the purely MA or random shock formulations, or the purely AR or inverted formats. Likewise, one can write the SARIMA or SARMA processes using any of the three formulations. The difference equation forms in which the SARIMA and SARMA processes are originally defined are given in [12.2.7] and [12.2.9], respectively. The random shock and inverted formulations are now described.

Random Shock Form

Because of both the invertibility and stationarity conditions referred to in Section 12.2.3, one can manipulate the operators algebraically and, hence, write the SARIMA process as either a pure MA process or a pure AR process. Using [12.2.7], the pure MA or *random shock form of the SARIMA model* is

$$z_t^{(\lambda)} = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)\nabla^d\nabla_s^D}a_t$$

$$
\begin{aligned}
&= a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + \cdots \\
&= a_t + \Psi_1 B a_t + \Psi_2 B^2 a_t + \cdots \\
&= (1 + \Psi_1 B + \Psi_2 B^2 + \cdots) a_t \\
&= \Psi(B) a_t
\end{aligned}
\qquad [12.2.15]
$$

where $\Psi(B) = 1 + \Psi_1 B + \Psi_2 B^2 + \cdots$, is the ***seasonal random shock or infinite MA operator,*** and Ψ_i is the ith parameter, coefficient or weight of $\Psi(B)$.

It is instructive to express a SARIMA process in the random shock in [12.2.15] for both theoretical and practical reasons. For example, the Ψ weights are needed to calculate the variance of forecasts when using a SARIMA model for forecasting. Also, one way to simulate data is to first write a SARIMA model in random shock form and then use this format of the model for producing synthetic data. Finally, by writing each member of a set of SARIMA models in random shock form, the models can be conveniently compared by examining the magnitudes and signs of the Ψ parameters.

In practice, one would first fit a SARIMA model to a given time series by following the model building approach described in Section 12.3 in order to obtain estimates for the AR and MA parameters. Subsequent to this, one may wish to calculate the Ψ_i parameters in [12.2.15], given the AR and MA parameters. The relationship for carrying out these calculations is contained in [12.2.15]. By definition, the following identity is true:

$$\Psi(B) = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)\nabla^d\nabla_s^D}$$

or

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D\Psi(B) = \theta(B)\Theta(B^s) \qquad [12.2.16]$$

By equating coefficients of B^k, $k = 1,2,\cdots$, on the left hand side of [12.2.16] to those on the right hand side, one can use the identity to solve for the Ψ_i coefficients in terms of the AR and MA parameters. Examples of performing these manipulations for nonseasonal ARMA models are presented in Section 3.4.3. Similar calculations can be carried out using [12.2.16] or a simplified version thereof for the seasonal case.

If no differencing operators are present and one is dealing with the SARMA model in [12.2.9] instead of the SARIMA model in [12.2.7], a similar procedure can be used to calculate the random shock weights. Simply remove the differencing operators from [12.2.16] and, once again, compare coefficients of B^k, $k = 0,1,2,\cdots$, to solve for the random shock parameters.

Another approach to solve for the random shock parameters for the SARIMA or SARMA model is to use [4.3.11] or [3.4.21] to solve for the Ψ_i's. This is accomplished by writing the SARIMA or SARMA model in unfactored form by using [12.2.12] or [12.2.11], respectively, and then making appropriate substitutions for the operators so that the nonseasonal formulae for calculating the seasonal random shock parameters can be used.

Inverted Form

To express the SARIMA model in inverted form, equation [12.2.7] can be written as

$$
\begin{aligned}
a_t &= \frac{\phi(B)\Phi(B^s)\nabla^d\nabla_s^D z_t^{(\lambda)}}{\theta(B)\Theta(B^s)} \\
&= z_t^{(\lambda)} - \Pi_1 z_{t-1}^{(\lambda)} - \Pi_2 z_{t-2}^{(\lambda)} - \cdots \\
&= z_t^{(\lambda)} - \Pi_1 B z_t^{(\lambda)} - \Pi_2 B^2 z_t^{(\lambda)} - \cdots \\
&= (1 - \Pi_1 B - \Pi_2 B^2 - \cdots) z_t^{(\lambda)} \\
&= \Pi(B) z_t^{(\lambda)} \qquad [12.2.17]
\end{aligned}
$$

where $\Pi(B) = 1 - \Pi_1 B - \Pi_2 B^2 - \cdots$, is the *seasonal inverted or infinite AR operator* and Π_i is the ith parameter, coefficient or weight of $\Pi(B)$. By comparing [12.2.17] and [12.2.16], one can see that

$$\Psi(B)^{-1} = \Pi(B) \qquad [12.2.18]$$

To calculate the inverted parameters, given that one knows the values of the AR and MA parameters, one can use the identity

$$\Pi(B) = \frac{\phi(B)\Phi(B^s)\nabla^d\nabla_s^D}{\theta(B)\Theta(B^s)}$$

or

$$\theta(B)\Theta(B^s)\Pi(B) = \phi(B)\Phi(B^s)\nabla^d\nabla_s^D \qquad [12.2.19]$$

which is obtained from [12.2.17]. To solve for the Π_i parameters, simply equate the coefficients of B^k for $k = 0,1,2,\cdots$, on the left hand side of [12.2.16] to those on the right hand side.

If one wishes to calculate the inverted parameters for the SARMA model in [12.2.9], simply eliminate the differencing operators in [12.2.19]. Following this, one can use [12.2.19] to determine the inverted parameters by equating coefficients of the B^k for $k = 0,1,2,\cdots$, on the left hand side to those on the right.

As was done for the random shock parameters, one can also use the nonseasonal formulae to determine the seasonal inverted parameters. Simply write the SARIMA or SARMA model in unfactored form as in [12.2.12] or [12.2.11], respectively, and then make appropriate substitutions into [4.3.14] or [3.4.27], respectively to solve for the seasonal inverted parameters.

12.3 MODEL CONSTRUCTION

12.3.1 Introduction

The most appropriate SARIMA model to fit to a given seasonal time series can be ascertained by following the identification, estimation, and diagnostic check stages of model construction. In the previous section, it is shown how the design of the multiplicative SARIMA family of models is a straightforward extension of the nonseasonal models presented in Part II of the book. Likewise, as is explained in this section, the tools used for SARIMA model building are either the same or else closely related versions of the nonseasonal model construction techniques discussed in Part III. The applications described in Section 12.4 clearly demonstrate that the model development techniques of this section can be conveniently and easily utilized for determining the most appropriate SARIMA or SARMA model to fit to a given time series. Finally, to allow the model building stages to be expeditiously and properly implemented in practice, one can employ a flexible decision support system such as the McLeod-Hipel Package described in Section 1.7.

12.3.2 Identification

Introduction

The purpose of the identification stage is to determine the nonseasonal and seasonal differencing required to produce stationarity and also the orders of both the nonseasonal and the seasonal AR and MA operators for the w_t series in [12.2.9]. Although each identification technique is discussed separately, in practical applications the output from all the techniques is interpreted and compared together in order to design the type of model to be estimated.

For some types of seasonal time series, it is known in advance whether or not the data sets should be transformed using the Box-Cox transformation in [12.2.1]. For instance, average monthly riverflow series often require a natural logarithmic transformation to cause the residuals of the fitted models to be approximately normally distributed and homoscedastic. In many applications, analysts may not realize that Box-Cox transformations are needed until after the model parameters have been estimated and the statistical properties of the residuals are examined. The analysts should keep in mind that usually a Box-Cox transformation does not change the design of the AR and MA operators needed in the model or models to fit to the transformed time series. However, this is not true in general, and as is pointed out by Granger and Newbold [1976], certain transformations can change the type of model to estimate for a given time series. Therefore, even though it is usually not necessary to perform the identification stage for the transformed data if it has already been done for the corresponding untransformed series, a practitioner should be aware that in some instances this may not be the case. When a transformation does change the type of model to be used, diagnostic checks would detect this fact and then the design of the model to fit to the transformed data can be properly identified.

For a SARIMA model application there should be at least seven years of seasonal data and also at least 50 data points overall in order to get reasonable MLE's (maximum likelihood estimates) for the model parameters. If one were analyzing a monthly series, one would require at least $12 \times 7 = 84$ observations. Therefore, one should proceed with the identification stage only if the minimum required amount of information is present.

Tools

When examining a specified time series analysis for the first time, one may wish to utilize the exploratory data analysis tools described in Section 22.3. The purpose of *exploratory data analysis* is to discover the basic statistical characteristics of a data set by examining simple graphical and numerical output. Subsequent to obtaining a general understanding of the statistical properties of the time series, one may wish to design a specific SARIMA model to fit to the series by studying the graphical output from the following techniques which are also used for designing nonseasonal models in Chapter 5:

1. **Plot of the Original Series** - A graph of the observations in the series against time is an important exploratory data analysis method that should *always* be used in model identification. Characteristics of the data which are usually easily uncovered from a perusal of a time series plot include seasonality, nonstationarity due to trends in the mean levels of the seasons or years, changing variance, extreme values, correlation or dependence among observations, and long term cycles. The nonstationarity present in the data can often be removed using seasonal and/or nonseasonal differencing. The seasonal and nonseasonal correlation in the time series can be modelled by appropriately deciding upon which AR and MA operators should be included in the SARIMA model. Graphs of the next four functions can be used for specifically designing the components needed in the SARIMA model.

2. **ACF (autocorrelation function)** - The theoretical ACF defined in [2.5.4] measures the amount of linear dependence between observations in a time series that are separated by k time lags. The sample estimate, r_k, for ρ_k, is given in [2.5.9] while approximate variances for r_k are given in [2.5.10] and [2.5.11]. To use the sample ACF in model identification, calculate and then plot r_k up to a maximum lag of roughly $\frac{\eta}{4}$ along with the approximate 95% confidence limits. The graph of the sample ACF and the other three graphs described below should include at least $2s$ or $3s$ lags, where s is the number of seasons per year. In this way, the cyclic behaviour caused by seasonality and any decaying or truncation properties of r_k over k, can be visually detected.

The first step is to examine a plot of the ACF to detect the presence of nonstationarity in the given series. For seasonal data with the seasonal length equal to s, the ACF often follows a wave pattern with peaks at s, $2s$, $3s$, and other integer multiples of s. As is shown by Box and Jenkins (1976, pp. 174-175), if the estimated ACF at lags that are integer multiples of the seasonal length s do not die out rapidly, this may indicate that seasonal differencing is needed to produce stationarity. Failure of other ACF estimates to damp out may imply that nonseasonal differencing is also required. If the length of the original series is η, the number of data points in the differenced series would be $\eta' = \eta - d + sD$. Li (1991) develops some statistical tests for determining the orders of differencing required for a seasonal time series.

If the stationary w_t series is not white noise, one can use the sample ACF to help decide upon which AR and MA parameters are needed in the SARIMA model. When the process is a pure MA$(0,d,q)\times(0,D,Q)_s$ model, the sample ACF truncates and is not significantly different from zero after lag $q+sQ$. For this case, the variance of r_k after lag $q+sQ$ is (Bartlett, 1946)

$$Var[r_k] \approx \frac{1}{\eta'}\left\{1+2\sum_{i=1}^{q+sQ} r_i^2\right\}, \quad k > q+sQ \qquad [12.3.1]$$

where η' stands for the length of the w_t series after differencing.

If r_k attenuates at lags that are multiples of s, this implies the presence of a seasonal AR component. The failure of the ACF to truncate at other lags may imply that a nonseasonal AR term is required.

As defined in [12.2.8], the stationary w_t series created by differencing either the original or transformed series is given as

$$w_t = \nabla^d \nabla_s^D z_t^{(\lambda)}$$

where the exponent λ indicates that the original z_t series may be transformed using the Box-Cox transformations in [12.2.1]. After the data have been differenced just enough times to produce both seasonal and nonseasonal stationarity, then check the ACF of the w_t series to determine the number of AR and MA parameters required in the model. The w_t series is also used at the other steps of the identification procedure. Of course, if no differencing is required, the w_t series is simply the $z_t^{(\lambda)}$ series. As noted earlier, the graph of the sample ACF for w_t should include at least $2s$ or $3s$ lags.

If a series is white noise, then r_k is approximately NID$(0, \frac{1}{\eta'})$. This result allows one to test whether a given series is white noise by checking to see if the ACF estimates are significantly different from zero. Simply plot confidence limits on the ACF diagram and see if a significant number of r_k values fall outside the chosen confidence interval.

3. **PACF (partial autocorrelation function)** - After writing the SARIMA or SARMA model in unfactored form as in [12.2.10] or [12.2.11], respectively, the theoretical PACF is defined for the w_t series in [3.2.17] using the Yule-Walker equations. Following the approach discussed in Section 3.2.2 and Appendix A3.1, the sample PACF can be estimated. For model identification, simply calculate and plot the sample PACF to at least lag $2s$ along with the 95% confidence limits which are calculated using [3.2.18], in which the length of the series is taken to be that of w_t. Employing rules put forward by authors such as Hipel et al. (1977), McLeod et al. (1977), and Hamilton and Watts (1978), the sample PACF can be utilized for deciding upon which AR and MA parameters are needed for properly representing the data.

   When the process is a pure AR$(p,d,0)\times(P,D,0)_s$ model, the sample PACF cuts off and is not significantly different from zero after lag $p + sP$. After lag $p + sP$, the sample PACF is approximately NID$(0, \frac{1}{\eta'})$.

   If the sample PACF damps out at lags that are multiples of s, this suggests the incorporation of a seasonal MA component into the model. The failure of the sample PACF to truncate at other lags may imply that a nonseasonal MA term is required.

4 **IACF (inverse autocorrelation function)** - The theoretical IACF (Cleveland, 1972) is defined in Section 5.3.6 and a method for estimating the sample IACF along with approximate 95% confidence limits is given in the same section. Theoretically, the IACF of the w_t series is defined to be the ACF of the $(q,d,p)\times(Q,D,P)_s$ process that is written as

$$\theta(B)\Theta(B^s)w_t = \phi(B)\Phi(B^s)a_t \qquad [12.3.2]$$

The model in [12.3.2] is called the *dual model* while the SARIMA$(p,d,q)\times(P,D,Q)_s$ model in [12.2.7] or [12.2.9] is referred to as the *primal model* (McLeod, 1984).

As is the case for all four functions discussed under points 2 to 5, the sample IACF is plotted up to a lag of at least $2s$ or $3s$ or not more than $\frac{\eta'}{4}$. If the w_t series is white noise, the sample IACF is approximately NID$(0,\frac{1}{\eta'})$. For the case of white noise, the values of the sample IACF should not fall outside the 95% confidence limits of $\pm\frac{1.96}{\sqrt{\eta'}}$ more than once in twenty lags.

For a pure AR$(p,d,0)\times(P,D,0)_s$ process, the sample IACF truncates and is not significantly different from zero after lag $p + sP$. If the sample IACF damps out but is still significant at lags s, $2s$, $3s$, etc., a seasonal MA component may be needed in the model. An additional nonseasonal MA component will cause the sample IACF to damp out for values between 0 and s, s and $2s$, etc., where decreasing but prominent peaks occur at s, $2s$, $3s$, etc., due to the seasonal MA term.

5. **IPACF (inverse partial autocorrelation function)** - The theoretical IPACF originally defined by Hipel et al. (1977) is presented in Section 5.3.7. The PACF for a SARMA model is by definition the IPACF of the dual model in [12.3.2]. In addition, a method for estimating the IPACF and obtaining approximate 95% confidence limits is explained in Section 5.3.7.

   For model identification, the sample IPACF and its 95% confidence limits are plotted up to a lag of at least $2s$ or $3s$. If the w_t series is white noise, then the values of the sample IPACF should not be significantly different from zero and should fall within the 95% confidence limits.

   For a pure MA$(0,d,q)\times(0,D,Q)_s$ model, the sample IPACF truncates and is not significantly different from zero after lag $q + sQ$. After lag $q + sQ$, the sample IPACF is approximately NID$(0,\frac{1}{\eta'})$. If the sample IPACF attenuates at lags that are multiples of s, this may indicate the presence of a seasonal AR component. When the IPACF fails to cut off at other lags, this implies the need for a nonseasonal AR term.

6. **Cumulative periodogram white noise test** - As was mentioned previously, the sample ACF plot is an accepted means of checking whether the given data are white noise. The sample PACF, IACF, and IPACF can also be employed in this capacity. However, the cumulative periodogram defined in [2.6.2] provides another means of checking for white noise.

   In addition to verifying whether a series is uncorrelated, the cumulative periodogram can also detects certain types of correlation. In particular, it is an effective procedure for finding hidden periodicities.

Summary

A plot of the original data portrays an overall view of how the time series is generally behaving and whether or not differencing is required. However, the sample ACF, PACF, IACF, and IPACF transform the given information into a format whereby it is possible to detect the number of AR and MA terms required in the model. In general, the ACF and the IPACF truncate for pure MA processes, while the PACF and IACF cut off for AR models. For mixed processes, all four functions attenuate. This behaviour of identification graphs for SARIMA models is summarized in Table 12.3.1.

The ACF and the IPACF possess similar general properties, while the PACF and the IACF have common attributes. However, the four functions are defined differently, and none of them behave exactly in the same fashion. In practice, the authors have found that if the PACF fails to detect a certain property of the time series, then the IACF often may be more sensitive and thereby may clearly display the presence of that property and vice versa. A similar situation exists between the ACF and the IPACF. In actual applications, it is necessary to consider simultaneously the output from all the functions in order to ascertain which model to estimate.

The incorporation of the IACF and the IPACF into the identification stage simplifies and substantiates model design because it is easier and more accurate to determine the proper SARIMA model to estimate. It is recommended that all of the identification plots be programmed for instantaneous display on a computer terminal screen. In this way, the identification stage can usually be completed in just a few minutes. The capability of making an immediate copy of any results portrayed on a screen provides a convenient method of keeping a permanent record. The decision support system for time series modelling described in Section 1.7 can be employed in this manner.

In Appendix A12.1, an alternative procedure for using the ACF to identify the parameters required in a SARMA model is given. This novel approach utilizes the structure of the multiplicative SARMA model by splitting the analysis using the ACF into nonseasonal and seasonal components.

Table 12.3.1. Behaviour of identification functions for SARIMA models.

	TYPES OF MODELS		
FUNCTION	Pure AR $(p,d,0)\times(P,D,0)_s$	Pure MA $(0,d,q)\times(0,D,Q)_s$	Mixed $(p,d,q)\times(P,D,Q)_s$
ACF	Attenuates	Truncates after lag $q+sQ$	Attenuates
PACF	Truncates after lag $p+sP$	Attenuates	Attenuates
IACF	Truncates after lag $p+sP$	Attenuates	Attenuates
IPACF	Attenuates	Truncates after lag $q+sQ$	Attenuates

12.3.3 Estimation

Introduction

Often, identification methods cannot clearly determine which is the single best SARIMA model to fit to the time series under study. Rather, anywhere from one to four model designs may be tentatively identified. At the estimation stage, *MLE*'s can then be obtained for the parameters in each of the models. Subsequently, discrimination methods can be used for selecting the best model from the set of calibrated models. The techniques for choosing the best model include the AIC discussed in this section as well as Section 6.3 and the diagnostic checks presented in Section 12.3.4. If none of the fitted models adequately describes the data, appropriate design modifications can be made before estimating the parameters of the most recent iteration and repeating the above procedure until a suitable model is found.

Algorithms

Two algorithms are discussed in this section for obtaining approximate MLE's for the parameters of the SARMA model fitted to the w_t series in [12.2.9]. Besides using the basic definition of the SARMA model, both methods are based on the assumption that the innovations are normally independently distributed with a mean of zero and variance of σ_a^2 [i.e. NID$(0,\sigma_a^2)$]. The first approach is to use the algorithm developed for the nonseasonal ARMA model while the second one is to employ a more computationally efficient procedure which takes into account the specific mathematical structure of the SARMA model. Because the orders of nonseasonal and seasonal differencing operators required to produce the stationary w_t series in [12.2.9] are not estimated but selected based upon identification results, only the parameters included in the SARMA model for fitting to w_t have to be estimated. Consequently, in this section, parameter estimation is discussed in terms of the SARMA$(p,q)\times(P,Q)_s$ model in [12.2.9] rather than the SARIMA$(p,d,q)\times(P,D,Q)_s$ model in [12.2.7]. Furthermore, since the mean level of w_t is zero due to differencing, the mean level or trend is not included in the SARMA model in [12.2.9]. If it were necessary to estimate a mean or trend this could be accomplished using the estimation methods discussed in this section.

In Chapter 6, the *modified sum of squares algorithm of McLeod (1977)* is suggested for estimating the parameters of the ARMA(p,q) model in [3.4.4]. As explained in that chapter, compared to other competing estimation methods, the modified sum of squares approach is both computationally and statistically efficient. The main steps in the algorithm are described in Appendix A6.1. To use the modified sum of squares method in Chapter 6 to estimate the parameters of the SARMA(p,q) model in [12.2.9], the first step is to write the model in the unfactored form given in [12.2.11]. Because the unfactored model in [12.2.11] can be considered as a special case of the ARMA(p,q) model, the modified sum of squares method can be used to obtain approximate MLE's for the model parameters and residuals.

The first estimation method works well when the number of seasons per year is not more than 12. However, for bimonthly data and weekly series for which $s = 24$ and 52, respectively, the estimation becomes computationally inefficient. To overcome this problem, the *maximum likelihood approach of McLeod and Salas (1983)* can be used. This estimation method, which is based upon the modified sum of squares method of McLeod (1977), is designed according to the multiplicative structure of the AR and MA operators in the SARMA model. It works for yearly ($s = 1$), monthly ($s = 12$), weekly ($s = 52$), daily ($s = 365$) as well as any other types of seasonal series. The main steps in the McLeod-Salas algorithm are described in Appendix A12.2.

The MLE's for a SARMA or other kind of time series model are asymptotically normally distributed. Because a *SE (standard error)* is estimated for each of the estimated parameters using the information matrix, one can check if an estimate is significantly different from zero. If the estimate is significant at the 5% significance level, its absolute magnitude should be larger than 1.96 SE. Usually, it is advisable to drop parameters which are not significantly different from zero from the SARMA model and then to re-estimate the parameters of the simplified model and then check if this model provides an adequate fit.

Model Discrimination

Subsequent to estimating the model parameters for the models separately fitted to the time series under study, one can calculate the value of the AIC for each model in order to select the model which has the minimum AIC value. This procedure is referred to as *MAICE (minimum AIC estimation)* and is described in detail in Section 6.3. The flow chart in Figure 6.3.1 explains the ways in which MAICE can be used in model construction for application purposes. One approach is to carry out an exhaustive AIC study by fitting a large range of SARMA models to the time series and then picking the one having the minimum AIC. In the second main approach, the identification techniques of Section 12.3.2 are used to select a handful of models for which the parameters and AIC values are estimated. Once again, one selects the model having the minimum AIC value.

The general formula for the AIC is given in [6.3.1] as (Akaike, 1974)

$$AIC = -2\ln(ML) + 2k$$

where ML denotes the maximized value of the likelihood function and k is the number of independently adjusted parameters in the model. Approximate formulae can be devised for determining the AIC for a SARIMA model which contains differencing operators. Because the amount of data has been reduced from a total of η to $\eta' = \eta - d - sD$ points when there is both nonseasonal and seasonal differencing, this will certainly affect the first term on the right-hand side of [6.3.1]. Hence, the AIC for a SARIMA model can be roughly calculated as

$$AIC = \frac{\eta}{\eta'}(-2\ln(ML)) + 2k \qquad [12.3.3]$$

where the maximized log likelihood is obtained by optimizing the log likelihood function defined in [A12.2.1]. The total number of model parameters is $k = p + q + P + Q + 1$, where the unity term allows for the estimate of the variance of the model residuals. Usually, the mean of the differenced series can be assumed to be zero. However, if the mean of the differenced series is also estimated, k must be increased by unity. Also, k is increased by one if $\lambda \neq 1$.

Another alternative for developing an AIC formula for a SARIMA model is to alter both of the terms on the right-hand side of [6.3.1]. As argued by Ozaki (1977), an increase in the number of data points contributes to decreasing the penalty due to the number of parameters. When the data are differenced both nonseasonally and seasonally, the number of data points decreases from η to $\eta' = \eta - d - sD$. This effect can be incorporated into the AIC by writing the formula as

$$AIC = \frac{\eta}{\eta'}(-2\ln(ML) + 2k) \qquad [12.3.4]$$

12.3.4 Diagnostic Checks

Introduction

The three assumptions underlying the innovations, a_t, $t = 1,2,\ldots,\eta'$, of the SARMA model in [12.2.9] are that the disturbances are independent, homoscedastic (i.e. have constant variance) and normally distributed. To check the foregoing residual stipulations, the estimated innovations, $\hat{a}_t$'s, or model residuals are required. The estimates for the $\hat{a}_t$'s are automatically calculated at the estimation stage along with the MLE's and SE's for the SARMA model parameters (see Appendices A12.2, A6.1 and A6.2).

A data transformation cannot correct dependence of the residuals because the lack of independence indicates that the present model is inadequate. Rather, the identification and estimation stages must be repeated in order to determine a suitable model having different model parameters. If the less important assumptions of homoscedasticity and normality are violated, they can often be corrected by a Box-Cox transformation of the data.

The residual assumptions for the SARMA model are identical to those stipulated for the nonseasonal ARMA model in [3.4.4]. Consequently, all of the diagnostic checks presented for the nonseasonal ARMA model in Chapter 7, can be used with the SARMA model. Some of these diagnostic methods are briefly described below but for detailed accounts of these and other model checking methods, the reader can refer to Chapter 7.

Tests for Whiteness

If a calibrated SARMA model adequately describes a time series, the estimated innovations, $\hat{a}_t$'s, or residuals should be white, due to the independence assumption of the a_t's. To determine whether the residuals are white noise, the best procedure is to examine the *residual autocorrelation function (RACF)*. Because the distribution of the RACF which is shown in the theorem below is now known, sensitive testing techniques are available for checking the independence assumption of a_t.

The theorem for the RACF is developed as follows. The ACF, $r_k(\hat{a})$, of the calculated residuals can be determined by

$$r_k(\hat{a}) = \sum_{t=k+1}^{\eta'} \left[\hat{a}_t \hat{a}_{t-k} \Big/ \left(\sum_{i=1}^{\eta'} \hat{a}_i^2 \right) \right] \qquad [12.3.5]$$

Define the vector of the first L values of the RACF as

$$\mathbf{r}(\hat{a}) = [r_1(\hat{a}), r_2(\hat{a}), \ldots, r_L(\hat{a})]' \qquad [12.3.6]$$

Denote by $\Psi_k(\Phi)$ the coefficient of B^k in the Maclaurin series expansion of $[\Phi(B^s)]^{-1}$ in powers of B, and similarly define $\Psi_k(\phi)$, $\Psi_k(\Theta)$, and $\Psi_k(\theta)$. Then it can be proved for large samples (McLeod, 1978) that

$$\mathbf{r}(\hat{a}) \approx N[0,(1/\eta')\mathbf{U}] \qquad [12.3.7]$$

where $\mathbf{U} = \mathbf{1}_L - \mathbf{X}\mathbf{T}^{-1}\mathbf{X}$, $\mathbf{1}_L$ is the identity matrix, $\mathbf{I} \approx \mathbf{X}'\mathbf{X}$ is the large-sample information matrix, and $X = [\Psi_{i-js}(\Phi), \Psi_{i-j}(\phi), \Psi_{i-js}(\Theta), \Psi_{i-j}(\theta)]$ are the i,j entries in the four partitions of the $\mathbf{X}$ matrix. The dimensions of the matrices X, $\Psi_{i-js}(\Phi)$, $\Psi_{i-j}(\phi)$, $\Psi_{i-js}(\Theta)$, and $\Psi_{i-j}(\theta)$ are, respectively, $L\times(P+p+Q+q)$, $L\times P$, $L\times p$, $L\times Q$, and $L\times q$. Previously, Box and Pierce (1970) obtained this result for the nonseasonal AR case, but the theorem listed here is valid for nonseasonal ARMA, SARMA, transfer function-noise, and intervention models.

There are two useful applications of the RACF distribution theorem. A sensitive diagnostic check is to first plot the RACF along with the asymptotic significance intervals for the RACF that are obtained from the diagonal entries of the matrix $\left(\frac{1}{\eta'}\right)\mathbf{U}$. If some of the RACF values are significantly different from zero, this may mean that the present model is inadequate. The important values of the RACF to examine are those at the first couple of lags and also at lags that are integer multiples of s for a seasonal model.

If the current model is insufficient, one can use the information from the RACF plot to help design an improved model before returning to the earlier stages of model construction. For example, a significantly large value of the RACF at lag s may indicate that a seasonal MA parameter is needed in the SARMA model. Upon updating the model design, the new model can be calibrated and checked for the presence of any further weaknesses.

A second but less sensitive test is to calculate and to perform a significance test for the modified Portmanteau statistic U_L (Li and McLeod, 1981). If L is large enough so that the weights $\Psi_k(\Phi)$, $\Psi_k(\phi)$, $\Psi_k(\Theta)$, and $\Psi_k(\theta)$ have damped out, then

$$U_L = \eta' \sum_{k=1}^{L} r_k^2(\hat{a}) + \frac{L(L+1)}{2\eta'} \qquad [12.3.8]$$

where L can be given as value of $2s$ to $4s$ such that the maximum value of L is not more than about $\frac{\eta'}{4}$. The statistic U_L is χ^2 distributed on $(L - P - p - Q - q)$ degrees of freedom. A test of this hypothesis can be done for model adequacy by choosing a level of significance and then comparing the value of the calculated χ^2 to the actual χ^2 value from the tables. If the calculated value is greater, on the basis of the available data, the present model is inadequate and, consequently, appropriate design changes must be made.

In Section 7.3.3, three Portmanteau test statistics are defined for carrying out whiteness tests with nonseasonal ARIMA models. The Portmanteau statistic defined in [12.3.8] for use with SARIMA models is based upon the statistic given in [7.3.6] for employment with nonseasonal ARIMA models. The seasonal equivalent of the test statistic in [7.3.5] is

$$U_L = \eta'(\eta' + 2) \sum_{k=1}^{L} r_k^2(\hat{a})/(\eta' - k) \qquad [12.3.9]$$

This statistic is χ^2 distributed on $(L - P - p - Q - q)$ degrees of freedom.

Test for Periodic Correlation

When a SARIMA model is fitted to a seasonal hydrological series such as the average monthly flows for the Saugeen River plotted in Figure VI.1, it may not be able to describe the periodic or seasonal correlation that may be contained in the series. This is because the SARIMA model assumes that the correlation structure contained in the series is the same throughout the year. However, the correlation between riverflow values for July and August, for instance, may be quite different from the correlation between April and March. This fact is confirmed in Section 14.4 where a periodic autoregressive model is fitted to the average monthly Saugeen riverflows.

To test whether or not periodic correlation is contained in the residuals of a fitted SARIMA model, one can employ a statistical test presented by McLeod (1993). The periodic autocorrelation at lag k for season m may be written as

$$r_k^{(m)}(\hat{a}_{r,m}) = \frac{\sum_{r=1}^{n} \hat{a}_{r,m}\hat{a}_{r,m-k}}{\sqrt{\sum_{r=1}^{n} \hat{a}_{r,m}^2 \sum_{r=1}^{n} a_{r,m+k}^2}} \qquad [12.3.10]$$

where $\hat{a}_{r,m}$ is the estimated innovation or residual for the rth year and mth season, n is the number of years of seasonal data, and s is the number of seasons per year. Over the s seasons, the residual autocorrelations at lag one given by $r_1^{(m)}(\hat{a}_{r,m})$, $m = 1,2, \ldots, s$, are approximately jointly normally distributed with mean zero, diagonal covariance matrix, and $var(r_1^{(m)}(\hat{a}_{r,m})) = n^{-1}$. A diagnostic check for detecting periodic autocorrelation in the residuals of a fitted SARIMA model is given by

$$S = n \sum_{m=1}^{s} (r_1^{(m)}(\hat{a}_{r,m}))^2 \qquad [12.3.11]$$

which should be approximately χ^2 distributed on s degrees of freedom, if the model is adequate. When the calculated value for S is larger than that found in the tables for a given significance level, the calibrated model does not capture the periodic correlation.

Normality Tests

As pointed out in Section 7.4, many standard tests are available to check whether data are normally distributed. Additionally, the graph of the cumulative distribution of the residuals should appear as a straight line when plotted on normality paper if the residuals are normally distributed (Section 7.4.3). For instance, the residuals should not be significantly skewed or possess a significantly large kurtosis coefficient under the assumption that the residuals are normally distributed (Section 7.4.2).

Homoscedasticity Checks

Heteroscedasticity or changes in variance can arise in a number of different ways including:

1. the variance changes over time,
2. the magnitude of the variance is a function of the current level of the series.

Sensitive significance tests for checking for the presence of the above kinds of heteroscedasticity are described in detail in Section 7.5.

12.3.5 Summary

By following the identification, estimation and diagnostic checks stages of model construction, one can conveniently determine a reasonable SARMA or SARIMA model for describing a time series. These three construction stages are summarized in Figure 12.3.1. This approach follows the general model building procedure shown in Figure 6.3.2. The applications in the next stage demonstrate how easy it is to carry out this SARMA model building approach in practice.

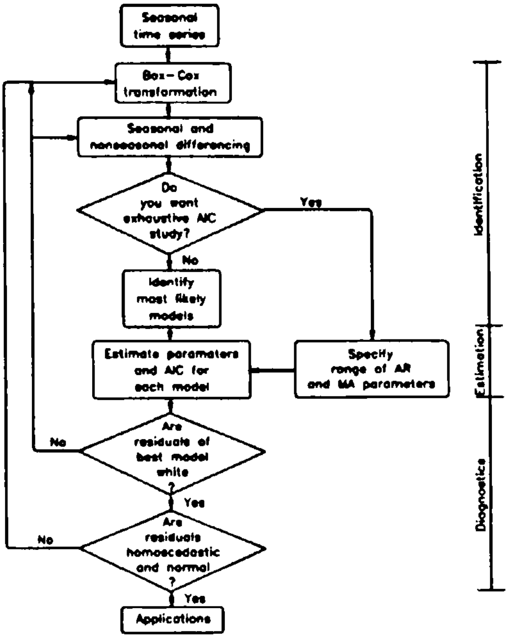

Seasonal time series
Box-Cox transformation
Seasonal and nonseasonal differencing
Do you want exhaustive AIC study?
Yes
No
Identify most likely models
Estimate parameters and AIC for each model
Specify range of AR and MA parameters
No
Are residuals of best model white ?
Yes
No
Are residuals homoscedastic and normal ?
Yes
Applications
Identification
Estimation
Diagnostics

Figure 12.3.1 Constructing a SARIMA model.

12.4 APPLICATIONS

12.4.1 Introduction

To demonstrate how the model construction procedures of Section 12.3 are conveniently utilized in practice, SARIMA models are fitted to three seasonal time series. In the first application, an appropriate SARIMA model is designed for describing the average monthly water consumption series of Figure VI.2 while in the second case study a SARIMA model is fitted to the average monthly concentrations of atmospheric CO_2 displayed in Figure VI.3.

As pointed out in the Foreword to Part VI, the seasonal series displayed in Figures VI.2 and VI.3 constitute data sets for which the level of the series within each series increases with time. This nonstationary characteristic is clearly depicted in Figures VI.2 and VI.3 by the increasing trend around which the seasonal data fluctuate in sinusoidal patterns. However, for the average monthly flows of the Saugeen River at Walkerton, Ontario, shown in Figure VI.1, the mean and variance within a particular season across the years are more or less stationary and, consequently, there is no upward trend. In the third application, the best SARIMA model to fit to this series is determined. However, as mentioned in the Foreword to Section VI, the most appropriate types of models to fit to the average monthly Saugeen Riverflows are the deseasonalized and periodic models of Chapters 13 and 14, respectively. Although a calibrated SARIMA model for the Saugeen flows could be used for forecasting, it cannot be employed for simulation. This is because the SARIMA model is not designed for preserving stationarity within each season.

12.4.2 Average Monthly Water Useage

The average monthly water consumption series in millions of litres per day is available from the Public Utilities Commission (1989) of London from 1966 to 1988. When fitting a SARIMA model to this series, the first step is to obtain appropriate exploratory data analysis and identification graphs referred to in Section 12.3.2 and then to compare the information displayed on these graphs in order to design the SARIMA model.

As explained in Section 22.3, one of the most informative exploratory data analysis tools is simply a plot of the given series against time. A plot of the monthly water consumption versus time in Figure VI.2 certainly reveals important characteristics about the observations. As also noted in the foreword to part VI as well as the introduction to this section, the sinusoidal curve in Figure VI.2 indicates that the data are seasonal. Indeed, the demand for water is highest during the warmer summer months and lowest during the winter time. Moreover, the increasing linear trend component indicates that the data in each month of the year are increasing with time. In some instances a physical understanding of the phenomenon being analyzed allows for the incorporation of deterministic components into the model to account for seasonality and/or trends. For instance, seasonality may be modelled by a Fourier series, while trend might be accounted for by a polynomial. However, for the water demand data, a purely stochastic SARIMA model is fit to the data. Following the explanation of Section 4.6, the nonseasonal and seasonal differencing take care of the stochastic trend while the AR and MA parameters can describe the remaining dependence among the observations. Consequently, the SARIMA model stochastically accounts for the inherent properties of the data.

From the graph of the water demand series in Figure VI.2, it is not obvious that a data transformation is required. However, the variance appears to be increasing in the last two years of the water demand series. When a Box-Cox transformation is estimated for the SARIMA model identified below, the best transformation is found to be $\lambda = -0.75$.

The upward trend in Figure VI.2 points out that seasonal and perhaps also nonseasonal differencing are needed for removing the nonstationary behaviour. The nonstationarity of the raw data is also confirmed by the fact that the sample ACF plotted in Figure 12.4.1 dies off very slowly. The seasonality of the water demand data is reflected in the attenuating sine wave pattern in this figure.

Differencing the data once seasonally removes the nonstationarity contained in the original series. A graph of the seasonally differenced series in Figure 12.4.2 shows that differencing has eliminated the linear trend component as well as the sine wave. Also notice that the seasonally differenced series in Figure 12.4.2 has 12 fewer data points than the graph in Figure VI.2 because of the monthly differencing. In order to compare conveniently the original water consumption series in Figure VI.2 to the seasonally differenced version of the series in Figure 12.4.2, the two series can be plotted on a single graph. Figure 12.4.3 displays the bivariate trace plot for these graphs which shows Figure VI.2 as the lower graph and Figure 12.4.2 as the upper plot. To avoid clutter and permit easier interpretation of the results, the ordinate axis is omitted. One can clearly see from Figure 12.4.3 how seasonal differencing has removed trend and sinusoidal components.

The sample ACF, PACF, IACF and IPACF are displayed in Figures 12.4.4 to 12.4.7 for the seasonally differenced water demand series. Because none of these graphs attenuate slowly, the seasonally differenced data of Figures 12.4.2 and 12.4.3 (upper plot) are stationary. When

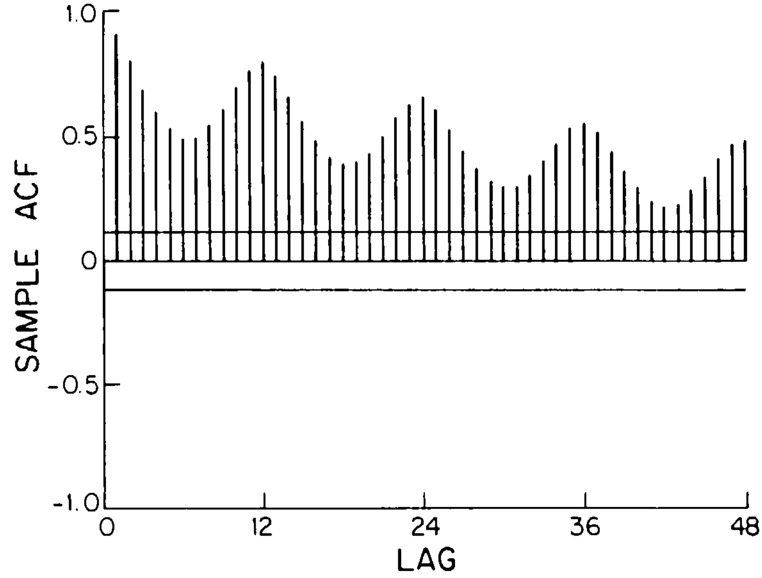


Figure 12.4.1. Sample ACF and 95% confidence limits for the average monthly water useage series from January, 1966, to December, 1988, for London, Ontario, Canada.

calculating the sample ACF in Figure 12.4.4, or, in general, when computing the ACF of any series that has been differenced, the mean of the w_t series in [12.2.8] is not removed. This procedure precludes missing any deterministic component that may still be present even after differencing.

In order to identify the number of AR and MA terms required in the model of the seasonally differenced monthly water useage data, the graphs of the sample ACF, the PACF, the IACF, and the IPACF that are shown in Figures 12.4.4 to 12.4.7, respectively, are interpreted simultaneously keeping in mind the main identification rules summarized in Table 12.3.1. Notice that both the sample ACF and IPACF have a significantly large value at lag 12. Moreover, because the sample PACF and IACF possess values that are decreasing in absolute magnitude at lags 12, 24, 36, and 48 (i.e. lags that are positive integer multiples of 12), this indicates the need for a seasonal MA term in the model.

Overall, the four identification graphs in Figures 12.4.4 to 12.4.7 complement one another in clearly pointing out the need for a seasonal MA parameter to include in the SARIMA model to fit to the seasonally differenced monthly water demand series. These graphs are also utilized for ascertaining which nonseasonal AR and MA parameters are needed. Because both the sample ACF and IPACF appear to die off across the first three or four lags, a nonseasonal AR parameter is required. Although the pattern is not strong, one could also interpret the sample PACF as attenuating during the first few lags. This indicates that a nonseasonal MA parameter may be needed in the model. Finally, notice that the sample IACF seems to cut off after the first

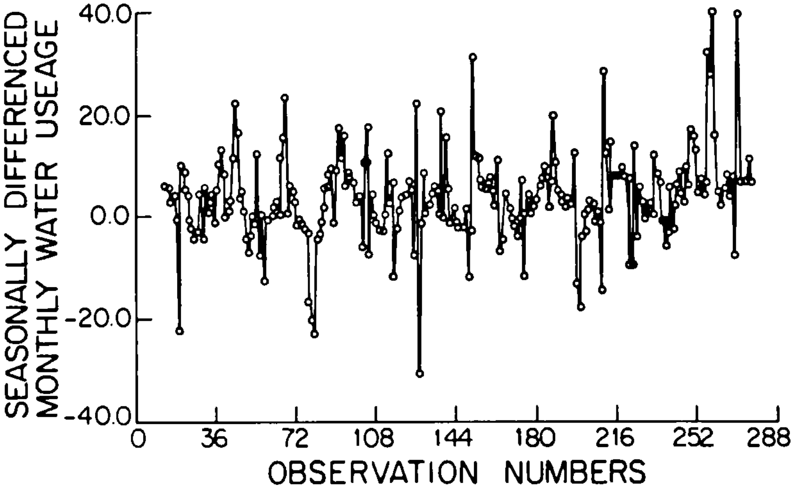


Figure 12.4.2. Graph of the seasonally differenced average monthly water useage series for London, Ontario, Canada.

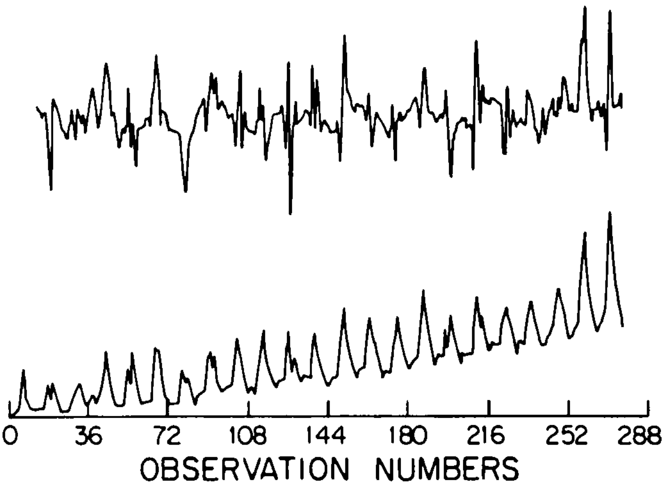


Figure 12.4.3. Bivariate trace plot of the average monthly water useage series (lower graph) for London, Ontario, Canada, from January, 1966, to December, 1988, and also the seasonally differenced water useage series (upper graph).

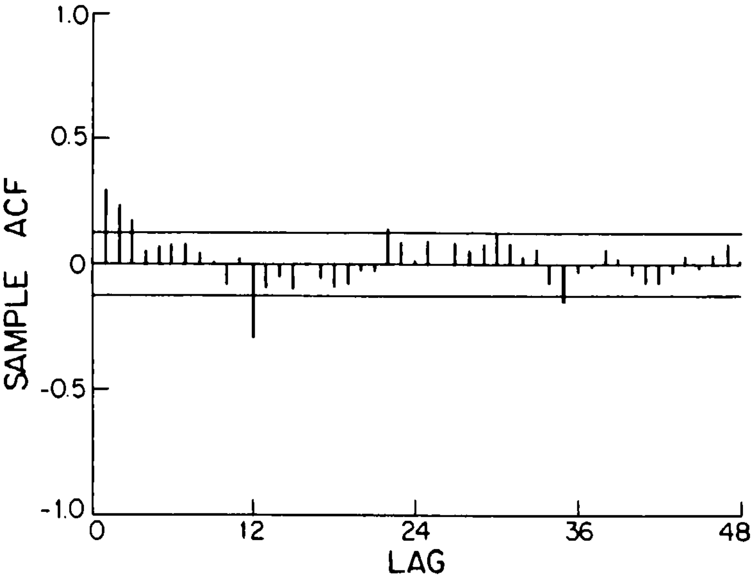


Figure 12.4.4. Sample ACF and 95% confidence limits for the seasonally differenced average monthly water useage series for London, Ontario, Canada.

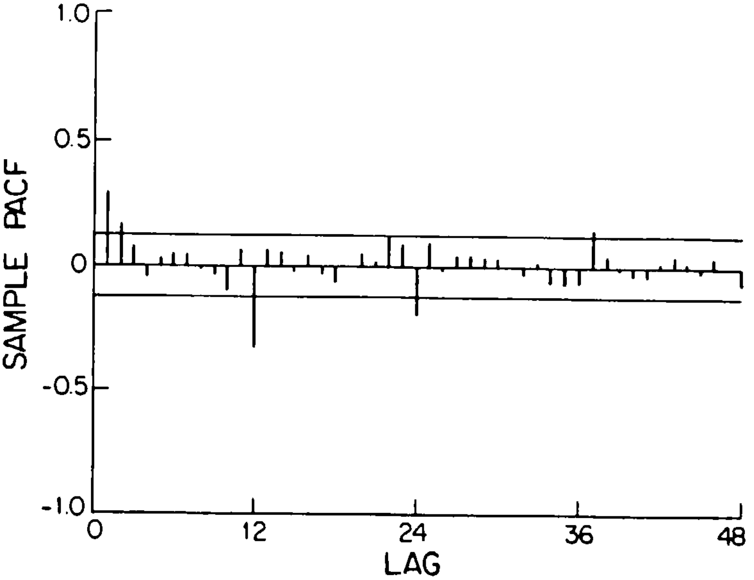


Figure 12.4.5. Sample PACF and 95% confidence limits for the seasonally differenced average monthly water useage series for London, Ontario, Canada.

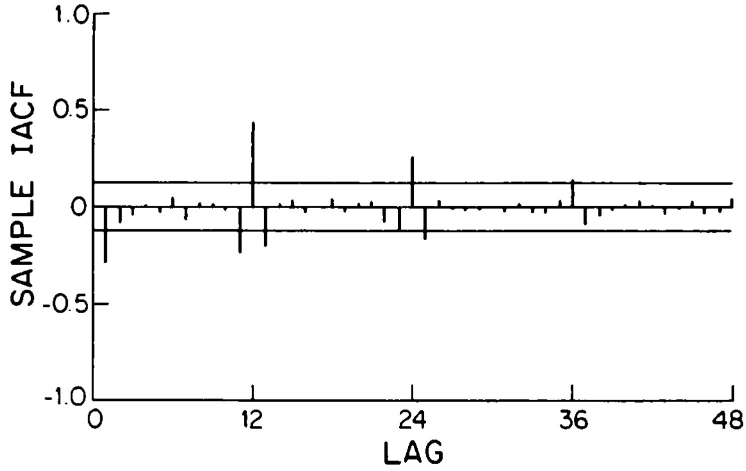


Figure 12.4.6. Sample IACF and 95% confidence limits for the seasonally differenced average monthly water useage series for London, Ontario, Canada.

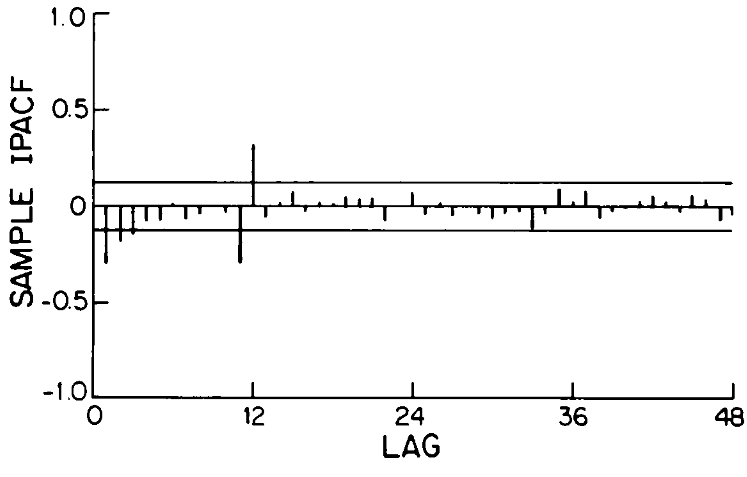


Figure 12.4.7. Sample IPACF and 95% confidence limits for the seasonally differenced average monthly water useage series for London, Ontario, Canada.

lag and, therefore, does not confirm the need for a nonseasonal MA parameter.

In summary, the identification plots given in Figures 12.4.4 to 12.4.7 indicate that an ARIMA$(1,0,1)\times(0,1,1)_{12}$ or an ARIMA$(1,0,0)\times(0,1,1)_{12}$ model should be fit to the water demand series. As mentioned earlier, a Box-Cox power transformation with $\lambda = -0.75$ is also required. When SARIMA models are calibrated for the seasonally differenced transformed water demand series, the SARIMA model possessing the lowest AIC calculated using [12.3.3] is the SARIMA$(1,0,1)\times(0,1,1)_{12}$ model. Table 12.4.1 lists the parameter estimates and the SE's for this model while [12.4.1] gives the corresponding difference equation.

$$(1-0.618B)(1-B^{12})z_t^{(\lambda)} = (1-0.297B)(1-0.851B^{12})a_t \qquad [12.4.1]$$

where the Box-Cox power parameter $\lambda = -0.75$ for water demand series, z_t.

Table 12.4.1. Parameter estimates and SE's for the SARIMA$(1,0,1)\times(0,1,1)_{12}$ model with $\lambda = -0.75$ fitted to the water consumption series for London, Ontario, Canada.

Parameters	MLE's	SE's
ϕ_1	0.618	0.123
θ_1	0.297	0.150
Θ_1	0.851	0.032
σ_a^2	1.86×10^{-6}	

Notice in Table 12.4.1 that the estimate for the nonseasonal MA parameter is about twice its SE and is, therefore, just barely significantly different from zero. On the other hand, the parameter estimates for ϕ_1 and Θ_1 are many times larger than their corresponding SE's. Recall that the need for incorporating both ϕ_1 and Θ_1 into the model are clearly indicated by the identification graphs.

The calibrated SARIMA model in [12.4.1] passes diagnostic checks for whiteness, normality and homoscedasticity refereed to in Section 12.3.4. Figure 12.4.8, for example, of the RACF for the fitted SARIMA model shows that the model residuals are white. Except for lag 22, all of the RACF values fall within the 95% confidence interval. This large value at lag 22 is probably due to chance alone and could not be removed by including other model parameters. One would expect that there is one chance in twenty that a value could fall outside the 95% confidence limits. Additionally, this large value does not occur at crucial lags such as 1, 12, 24, and 36. Other diagnostic checks reveal that both the homoscedasticity and the normality assumption for the residuals are fulfilled. Therefore, on the basis of the information used, the chosen SARIMA model [12.4.1] adequately models the monthly water useage data.

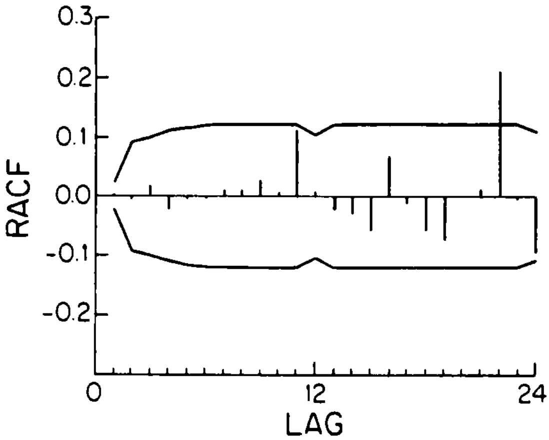


Figure 12.4.8. RACF and 95% confidence limits for the SARIMA(1,0,1)×(0,1,1)$_{12}$ model with $\lambda = -0.75$ fitted to the average monthly water useage series for London, Ontario, Canada.

12.4.3 Average Monthly Atmospheric Carbon Dioxide

Bacastow and Keeling (1981) as well as Keeling et al. (1982) list average monthly concentrations of atmospheric CO_2 measured at the Mauna Loa Observatory on the Island of Hawaii. Figure VI.3 displays a plot of these observations from January, 1965, to December, 1980. Because carbon dioxide is a principal greenhouse gas that could cause global warming, the monitoring and analysis of CO_2 data such as the series in Figure VI.3 is of wide interest to environmental scientists, political decision makers and, indeed, the general public. As pointed out by Bacastow and Keeling (1981), the increase per year in carbon dioxide concentration in the atmosphere is one of the important observations for understanding the carbon cycle.

Blue (1991) reports upon the work of Wahlen et al. (1991) who are taking measurements of CO_2 in the air of bubbles in the GISP 2 (Greenland Ice Sheet Project 2) ice core using a dry extraction technique and tunable diode laser absorption spectroscopy. The CO_2 record spans the years from 1530 to 1940 and includes parts of the little ice age, a time of abnormally cold temperatures in Europe during the 16*th* and 18*th* centuries. Wahlen et al. (1991) have found that there were no significant changes in CO_2 concentrations during the little ice age. However, since about 1810, the CO_2 concentrations have started to increase due to industrialization and other related land use changes such as deforestation and urbanization. Moreover, the date of the onset of increasing CO_2 in about 1810 in the GISP 2 ice core is similar to that discovered in the Siple core from Antarctica (Neftel et al., 1982) and also in another Antarctic ice core analyzed by Pearman et al. (1986). Finally, the data retrieved from the Greenland ice core GISP 2 are consistent with observations from Mauna Loa.

Figure VI.3 clearly depicts an increasing linear trend in the monthly CO_2 levels over the years. The sinusoidal curve wrapped around the trend demonstrates that the data are sinusoidal with larger values occurring in May or June of each year. Because of the nonstationarity in Figure VI.3, differencing is required to remove trends and hence create a stationary series. Figure 12.4.9 shows a bivariate trace plot for which the original CO_2 series of Figure VI.3 is given in the lower half of the graph and the stationary series is plotted at the top when the given series is differenced both seasonally (i.e. $D = 1$) and nonseasonally ($d = 1$). Notice from the top series the way in which differencing has removed the increasing trend as well as the distinct seasonality shown in the original lower series. Also, because of the differencing, the upper series is 13 ($d + D = 1 + 12 = 13$) data points shorter than the original lower series.

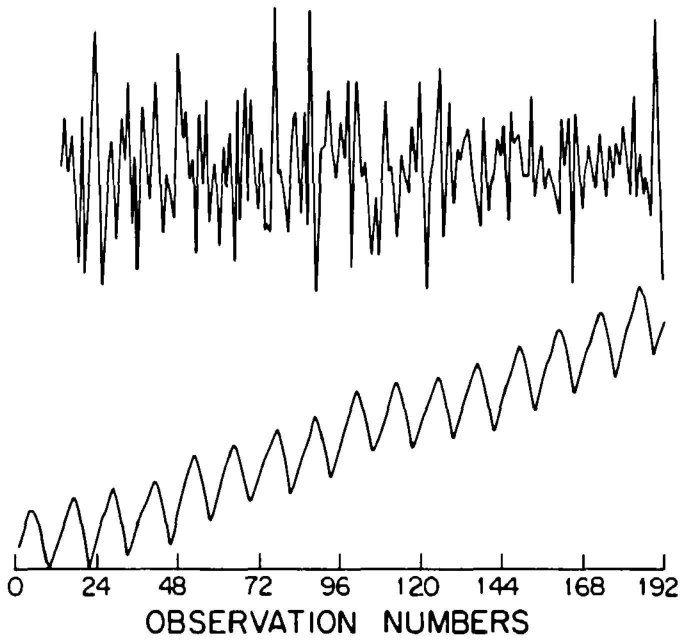


Figure 12.4.9. Bivariate trace plot of the average monthly concentrations of atmospheric CO_2 (mole fraction in ppm) (lower graph) measured at Mauna Loa Observatory in Hawaii from January, 1965, to December, 1980, and also the nonseasonally and seasonally differenced CO_2 series (upper graph).

To discover the AR and MA parameters required to model the top series in Figure 12.4.9, one can simultaneously examine the sample ACF, PACF, IACF and IPACF shown in Figures 12.4.10 to 12.4.13, respectively. The significantly large values of the sample ACF and IPACF at lag 12 means that a seasonal MA parameter is needed in the model. This finding is also confirmed by the fact that the sample PACF and IACF attenuate at lags that are positive integer multiples of 12 (i.e. lags 12, 24, 36 and 48). The slightly large values at lag one for the sample ACF points out that a nonseasonal MA parameter may be needed. The fact that the sample IACF possesses attenuating values after lags 1, 12, 24 and 36 may indicate that a nonseasonal MA parameter is required. Consequently, the most appropriate model to fit to the monthly CO_2 data is probably a SARIMA$(0,1,1)\times(0,1,1)_{12}$ model. When comparing other possible SARIMA models such as the SARIMA$(1,1,1)\times(0,1,1)_{12}$ and $(1,1,0)\times(0,1,1)_{12}$ to the aforementioned model, the AIC value calculated using [12.3.3] is lowest for the SARIMA$(0,1,1)\times(0,1,1)_{12}$ model.

Table 12.4.2 lists the MLE's and SE's for the parameters of the SARIMA$(0,1,1)\times(0,1,1)_{12}$ model fitted to the monthly CO_2 data set. In difference equation form, this calibrated SARIMA model is written as

$$(1-B)(1-B^{12})z_t = (1-0.336B)(1-0.831B^{12})a_t \qquad [12.4.2]$$

where z_t represents the average monthly CO_2 series value at time t.

The estimated model in [12.4.2] provides a reasonable fit to the CO_2 series according to diagnostic checks for whiteness, normality and constant variance described in Sections 7.3 to 7.5, respectively. Because all of the values for the RACF fall within the 95% confidence limits in Figure 12.4.14, the model residuals are not significantly correlated and, hence, are white. Moreover, the calibrated model performs reasonably well with respect to normality and homoscedasticity checks. For example, the value of the test statistic for changes in variance depending on the current level (see Section 7.5.2) is -2.232 with a SE of 1.296. Because the test statistic falls within two SE's of zero, one can argue that the statistic is not significantly different from zero and, hence, the residuals are homoscedastic.

12.4.4 Average Monthly Saugeen Riverflows

The average monthly flows for the Saugeen River at Walkerton, Ontario, Canada, are displayed in Figure VI.1. When modelling monthly riverflow series it is usually necessary to take natural logarithms of the data to alleviate problems with heteroscedasticity and/or nonnormality in the model residuals. Therefore, the Box-Cox parameter λ is set equal to zero in [12.2.1] for the Saugeen flows. Following this, an examination of the sample ACF for the logarithmic data reveals that the data has to be differenced once seasonally in order to remove seasonal nonstationarity. By studying the properties of the sample ACF, PACF, IACF and IPACF graphs for the seasonally differenced logarithmic Saugeen time series, it is found that the best design is a SARIMA$(1,0,1)\times(0,1,1)_{12}$ model. Diagnostic checks for the residuals of the calibrated model indicate that the model provides an adequate fit to the data. However, it is found that the periodic autocorrelation test statistic in [12.3.11] for the fitted SARIMA has a significantly large value of 59.6 on twelve degrees of freedom. Consequently, the SARIMA model residuals possess significant periodic correlation. Therefore, in Section 14.4 a periodic autoregressive model is fitted to the logarithmic average monthly Saugeen riverflows in order to

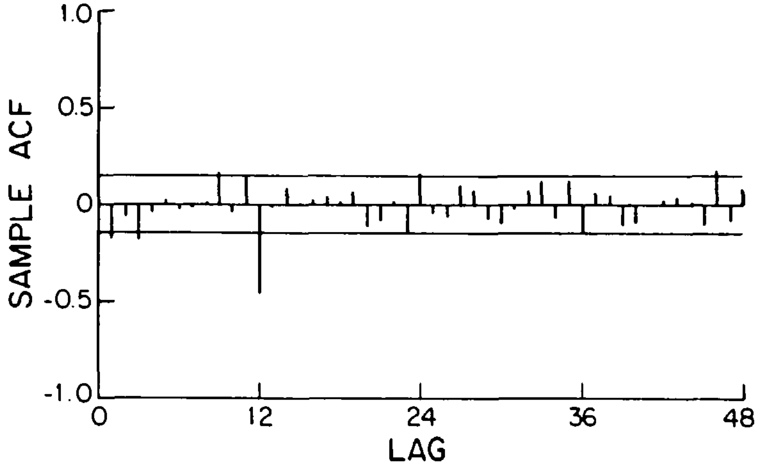


Figure 12.4.10. Sample ACF and 95% confidence limits for the differenced ($d = D = 1$) average monthly atmospheric CO_2 concentrations (mole fraction in ppm) measured at Mauna Loa Observatory in Hawaii.

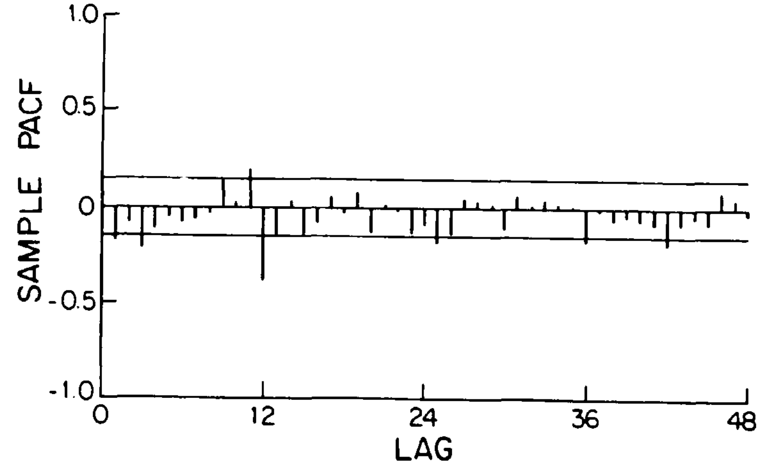


Figure 12.4.11. Sample PACF and 95% confidence limits for the differenced ($d = D = 1$) average monthly atmospheric CO_2 concentrations (mole fraction in ppm) measured at Mauna Loa Observatory in Hawaii.

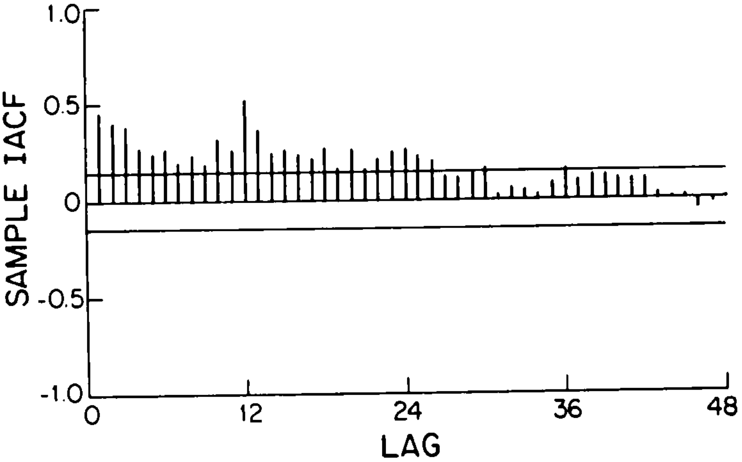


Figure 12.4.12. Sample IACF and 95% confidence limits for the differenced ($d = D = 1$) average monthly atmospheric CO_2 concentrations (mole fraction in ppm) measured at Mauna Loa Observatory in Hawaii.

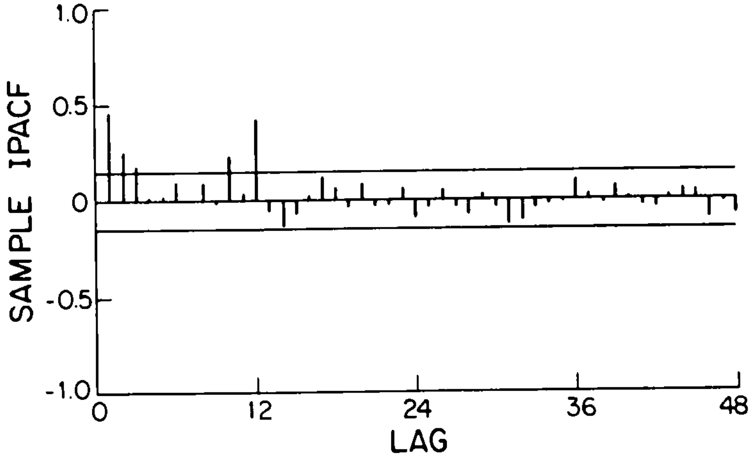


Figure 12.4.13. Sample IPACF and 95% confidence limits for the differenced ($d = D = 1$) average monthly atmospheric CO_2 concentrations (mole fraction in ppm) measured at Mauna Loa Observato:y in Hawaii.

Table 12.4.2. Parameter estimates and SE's for the SARIMA(0,1,1)×(0,1,1)$_{12}$ model fitted to the average monthly atmospheric CO_2 concentrations measured at Mauna Loa Observatory in Hawaii.

Parameters	MLE's	SE's
θ_1	0.336	0.070
Θ_1	0.831	0.041
σ_a^2	1.01×10^{-1}	

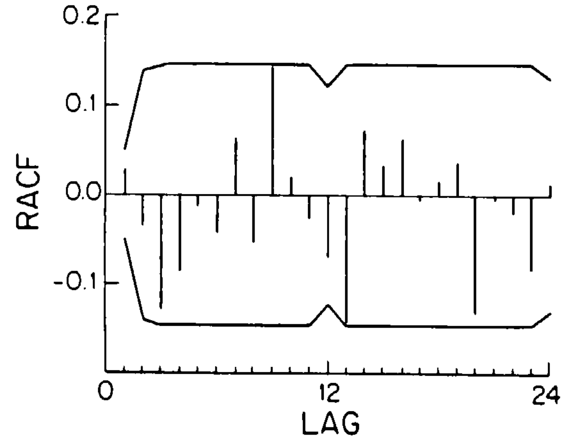


Figure 12.4.14. RACF and 95% confidence limits for the SARIMA(0,1,1)×(0,1,1)$_{12}$ model fitted to the average monthly atmospheric CO_2 concentrations (mole fraction in ppm) measured at Mauna Loa Observatory in Hawaii.

properly model the periodic correlation.

Equation [12.3.3] can be employed for calculating the AIC of the SARIMA model fitted to the Saugeen flows. The value of the AIC is found to be 3435.43. As shown in Chapters 13 and 14 for the values of the AIC determined for the deseasonalized and periodic models, respectively, the estimate of the AIC is much higher for the calibrated SARIMA model. This is because deseasonalized and periodic models are specifically designed for preserving certain kinds of stationarity within each season of the average monthly Saugeen riverflows.

12.5 FORECASTING AND SIMULATING WITH SARIMA MODELS

After fitting the SARIMA model in [12.2.7] to a given seasonal time series, the calibrated model can be employed for forecasting and simulation. In Chapter 8, minimum mean square error forecasts are defined and procedures are presented for calculating MMSE forecasts for nonseasonal ARMA (Chapter 3) and ARIMA (Chapter 4) models. Section 15.2.2 explains how MMSE forecasts are conveniently determined for a SARIMA model when the model is written using the generalized form in [12.2.12]. If the data have been transformed using a Box-Cox transformation, one can employ the procedure explained in Section 15.2.2 as well as Section 8.2.7 to determine the forecasts in the original untransformed domain.

In Section 12.4.2 of the previous section, the most appropriate SARIMA model to fit to the average monthly water useage series displayed in Figure VI.2 is found to be the SARIMA(1,0,1)×((0,1,1)$_{12}$ model with $\lambda=-0.75$ written in [12.4.1]. Figure 12.5.1 depicts the MMSE forecasts for the water demand series from January 1989 until December 1990 that are calculated for the fitted SARIMA model by following the procedure given in Section 15.2.2. The seasonal characteristics of these forecasts can be clearly seen as a sinusoidal pattern on the right hand side of Figure 12.5.1 that fall within their 90% probability limits.

The model for the SARIMA(0,1,1)×(0,1,1)$_{12}$ model fitted to the average monthly concentrations of atmospheric CO_2 is written in [12.4.2]. The MMSE forecasts for this model are displayed on the right hand side of Figure 12.5.2 along with the 90% probability limits. As is also the case in Figure 12.5.1, the forecasts follow the seasonal sinusoidal shape exhibited by the historical observations.

When simulating with a SARIMA model, the first step is to write the model in unfactored form as in [12.2.11]. Next, WASIM1 or WASIM2 explained in Sections 9.4 and 9.5, respectively, is employed to determine simulated data for the w_t series in [12.2.11]. Subsequently, the algorithm of Section 9.5.2 is utilized for integrating the systematic w_t values to obtain the simulated $z_t^{(\lambda)}$ series. Finally, the inverse Box-Cox transformation in [9.6.2] is invoked to procure the corresponding z_t synthetic data in the untransformed domain. Parameter uncertainty can be entertained by employing the WASIM3 algorithm of Section 9.7.

12.6 CONCLUSIONS

The SARIMA and SARMA models defined in [12.2.7] and [12.2.9], respectively, are designed for modelling time series which exhibit nonstationarity both within and across seasons. An inherent advantage of the SARIMA family of models is that relatively few model parameters are required for describing these types of time series. As demonstrated by the three applications in Section 12.4, the model construction techniques of Section 12.3 can be conveniently and expeditiously implemented in practice for designing, calibrating and checking SARIMA models. Many other applications of SARIMA models to water resources and environmental time series can be found in the literature. For instance, Irvine and Eberhardt (1992) fit SARIMA models to lake level time series.

As explained in Section 12.5 and also Section 15.2.2, a calibrated SARIMA model can be employed for forecasting and simulation. In Chapter 15, the results of forecasting experiments demonstrate that the deseasonalized and periodic models of Chapters 13 and 14, respectively, forecast better than SARIMA models when forecasting monthly riverflow time series. This is

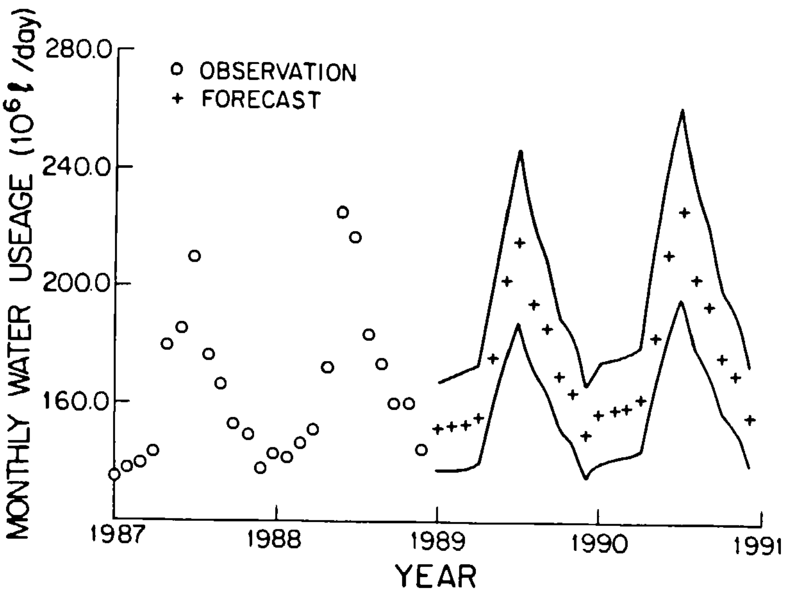


Figure 12.5.1. MMSE forecasts along with their 90% probability intervals for the SARIMA(1,0,1)x(0,1,1)$_{12}$ model with $\lambda = -0.75$ fitted to the average monthly water useage (million litres per day) series from 1966 to 1988 for London, Ontario, Canada.

because the SARIMA model is designed for modeling the type of series shown in Figures VI.2 and VI.3 rather than the one in Figure VI.1.

If there were more than one seasonal cycle in a series, other sets of seasonal operators could be incorporated into the SARIMA model to handle this situation. Furthermore, one could also easily design a seasonal FARMA (fractional ARMA model) by directly extending the definitions in Chapter 11 for nonseasonal FARMA models to the seasonal case.

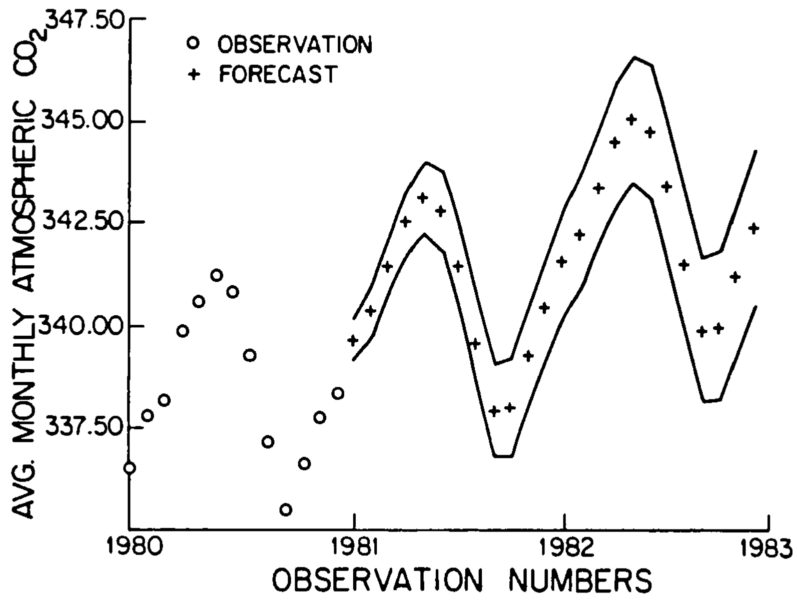


Figure 12.5.2. MMSE forecasts along with their 90% probability intervals for the SARIMA(0,1,1)×(0,1,1)$_{12}$ model fitted to the average monthly water concentrations of atmospheric CO_2 (mole fraction in ppm) measured at Mauna Loa observatory in Hawaii.

APPENDIX A12.1

DESIGNING

MULTIPLICATIVE SARIMA MODELS

USING THE ACF

The multiplicative SARMA$(p,q)\times(P,Q)_s$ model for fitting to the stationary w_t series in [12.2.8] is defined in [12.2.9]. The theoretical ACF of w_t is given by

$$\rho_k = \gamma_k/\gamma_0, \quad k = 0,1,2,\cdots$$

where $\gamma_k = E[w_t w_{t-k}]$. Using a result of Godolphin (1977, [3.4]), it follows that for s large enough

$$\rho_{i\pm js} \approx \rho_i(u_t)\rho_j(U_t) \quad \text{for } i = 0,1,\ldots,s/2 \text{ and } j = 0,1,\cdots, \qquad \text{[A12.1.1]}$$

where $\rho_{i\pm js}$ is the ACF of w_t at lag $i\pm js$ and $\rho_i(u_t)$ and $\rho_j(U_t)$ are the values of the ACF at lags i and j of u_t and U_t in the models $\phi(B)u_t = \theta(B)a_t$ and $\Phi(B^s)U_t = \Theta(B^s)a_t$, respectively. Godolphin (1977, [3.4]) demonstrates using a vector representation that [A12.1.1] holds exactly when $p = P = 0$ and $s > 2q$. As shown at the end of this appendix, this result may also be derived using the autocovariance generating function (Box and Jenkins, 1976, p. 81). In any case, the general approximation follows from the fact that a SARMA$(p,q)\times(P,Q)_s$ model can be approximated by a $(0,q)\times(P,Q)_s$ model for suitable Q when s is large enough.

A graphical interpretation, along similar lines to that suggested by Hamilton and Watts (1978) for the PACF may be given. First, the regular nonseasonal component pattern is defined as the ACF of u_t plus repetitions centered about the seasonal lags $0,s,2s,\cdots$. Figure A12.1.1 shows the regular component pattern corresponding to a $(1,0)(P,Q)_{12}$ model with $\phi_1 = 0.6$. Next, the seasonal component pattern is determined from the ACF of U_t as illustrated in Figure A12.1.2 for the case of a $(p,q)(1,0)_{12}$ model with $\Phi_1 = 0.6$. Then, the approximation to the ACF of w_t is the product of the regular and seasonal component patterns. Figure A12.1.3 shows the resulting approximation for the $(1,0)(1,0)_{12}$ model with $\phi_1 = \Phi_1 = 0.6$.

In conclusion, the general interpretation presented in this appendix is useful for seasonal model identification and does not appear to have been pointed out previously. For some applications, the employment of this procedure can simplify the identification process.

Generating Function Proof of Godolphin's Result

Theorem: If $s > 2q$, the autocovariance function, γ_k, of a $(0,q)\times(P,Q)_s$ model may be expressed for nonnegative lags as,

$$\gamma_{js\pm i} = \begin{cases} \gamma_i(u_t)\gamma_j(U_t) & \text{for } 0 \le i \le q \text{ and } 0 \le j \le Q, \\ 0, & \text{otherwise,} \end{cases}$$

where $\gamma_i(u_t)$ and $\gamma_j(U_t)$ denote the autocovariance functions of the processes $\phi(B)u_t = \theta(B)a_t$ and $\Phi(B)U_t = \Theta(B)a_t$.

Proof: For convenience, it may be assumed that $var(a_t) = 1$.

$$\begin{aligned}\Gamma(B) &= \theta(B)\theta(B^{-1})/[\Phi(B)\Phi(B^{-1})] \\ &= \sum_{k=-\infty}^{\infty} \Gamma_k B^k\end{aligned}$$

Then the autocovariance generating function of the $(0,q)(P,Q)_s$ model may be written as

$$\begin{aligned}\gamma(B) &= \sum_{k=-\infty}^{\infty} \gamma_k B^k \\ &= \theta(B)\theta(B^{-1})\Gamma(B^s)\end{aligned}$$

$$= \left[\left(\sum_{l=0}^{q} \sum_{i=0}^{q-l} \theta_i \theta_{i+l} \right) (B^l + B^{-l}) \right] \Gamma(B^s),$$

where $\theta_0 = -1$. Thus, provided $s > 2q$, the coefficient of B^{js+i} $(i = 0, \cdots ,q)$ is

$$\gamma_{js+i} = \Gamma_j \sum_{l=0}^{q-i} \theta_l \theta_{l+i}$$

Similarly, the coefficients of B^{js+i} and B^{js-i} can be shown to be equal.

APPENDIX A12.2

MAXIMUM LIKELIHOOD ESTIMATION

FOR

SARMA MODELS

McLeod and Salas (1983) provide an algorithm for calculating an approximation to the likelihood function of the multiplicative SARMA model in [12.2.9]. Their algorithm specifically takes advantage of the multiplicative structure of the nonseasonal and seasonal AR and MA operators in the SARMA model. The conditional, unconditional or iterated unconditional method of Box and Jenkins (1976) may be used in the algorithm of McLeod and Salas (1983) in conjunction with an approximation to the determinant term (see McLeod (1977) and also Appendix 6.1) to obtain an accurate and highly efficient algorithm. In fact, McLeod and Salas (1983) point out that other competing algorithms become computationally infeasible when the seasonal period s becomes much larger than 12, as in the cases of half-monthly ($s = 24$), weekly ($s = 52$) or daily ($s = 365$) time series.

In this appendix, the theory behind the algorithm of McLeod and Salas (1983) is outlined. The reader can refer to the paper of McLeod and Salas (1983) for a more detailed description of the theory and method of application as well as a listing of the Fortran computer program for the algorithm.

The SARMA $(p,q)\times(P,Q)_s$ model for fitting to a series w_t of length $\eta' = \eta - d - sD$ is defined in [12.2.9]. Let the vector of model parameters be given by

$$\beta = (\phi_1,\phi_2, \ldots ,\phi_p,\theta_1,\theta_2, \ldots ,\theta_p,\Phi_1,\Phi_2, \ldots ,\Phi_p,\Theta_1,\Theta_2, \ldots ,\Theta_Q)$$

Although the SARMA model may be considered as a special case of the ARMA (p^*,q^*) model in [12.2.14] by taking $p^* = sP$, $q^* = q + sQ$, $\phi^*(B) = \Phi(B^s)\phi(B)$ and $\theta^*(B) = \Theta(B^s)\theta(B)$, a more efficient estimation algorithm can be developed utilizing the multiplicative structure of the SARMA model.

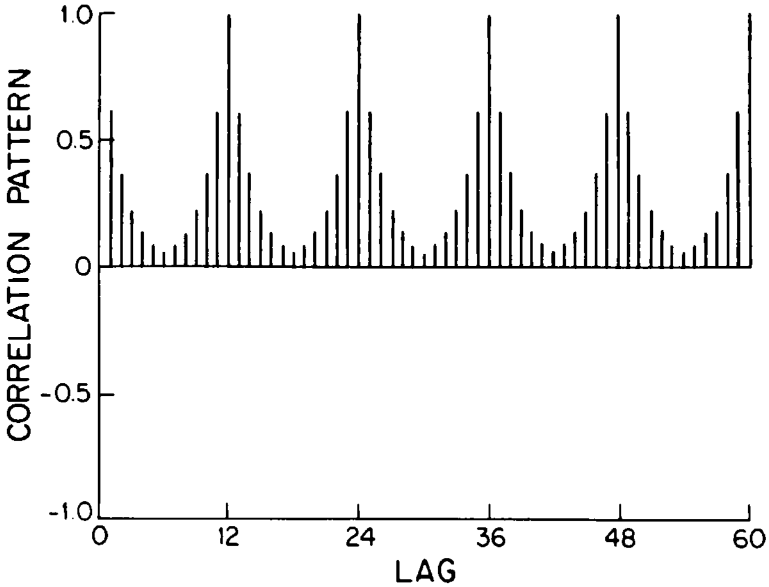


Figure A12.1.1. A regular nonseasonal component pattern of a SARIMA(1,0)$(P,Q)_{12}$ model with $\phi_1 = 0.6$.

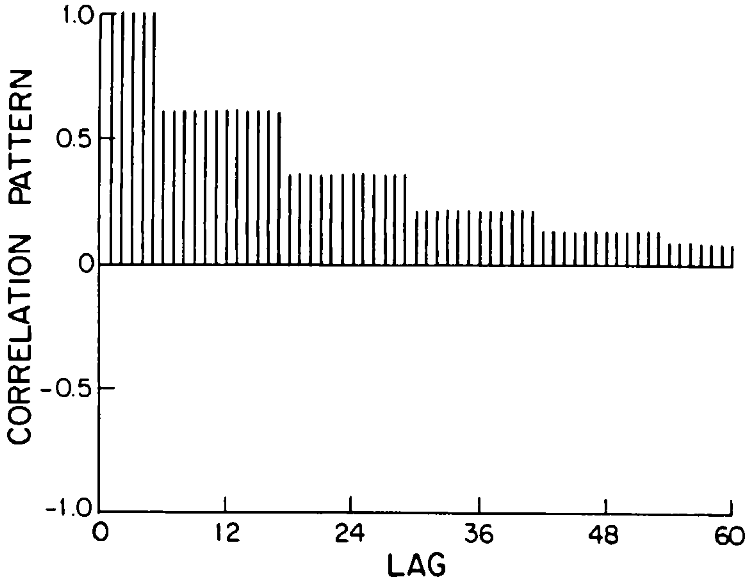


Figure A12.1.2. Seasonal component pattern of a SARMA$(p,q)\times(1,0)_{12}$ with $\Phi_1 = 0.6$.

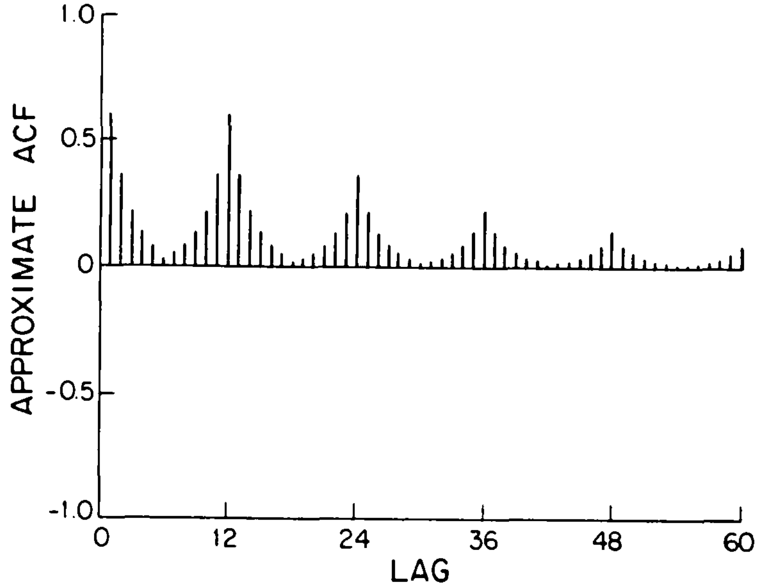


Figure A12.1.3. Approximate ACF of a SARMA$(1,0)(1,0)_{12}$ model with $\phi_1 = 0.6$ and $\Phi_1 = 0.6$.

Given the observations w_t, $t=1,2,\cdots,\eta'$, the exact log-likelihood function maximized over the variance, σ_a^2, of the innovation, a_t, may be written, apart from an arbitrary constant,

$$\log L(\beta) = -\eta'\log(S_m/\eta')/2 \qquad [A12.2.1]$$

where S_m, the modified sum of squares is

$$S_m = S[M_{\eta'}(p,q,P,Q,s)]^{-1/\eta'}. \qquad [A12.2.2]$$

S represents the unconditional sum of squares of Box and Jenkins (1976) defined by

$$S = \sum_{t=-\infty}^{n} [a_t]^2, \qquad [A12.2.3]$$

where $[.]$ denotes expectation given $w_1, w_2, \cdots, w_{\eta'}$.

The evaluation of S by the iterative unconditional sum of squares method may involve two types of truncation error. First, the infinite sum in [A12.2.3] is replaced by

$$S = \sum_{t=1-T}^{n} [a_t]^2 \qquad [A12.2.4]$$

for suitably large T. Theoretically, T should be chosen so that

$$\gamma_0/\sigma_a^2 - \sum_{i=0}^{T}\psi_i^2 < e_{t01} \qquad [A12.2.5]$$

where $\gamma_0 = var(w_t)$, ψ_i is the coefficient of a_{t-i} in the infinite MA representation of [12.2.9] and e_{t01} is an error tolerance. Thus, if the model contains a AR factor with roots near the unit circle, a fairly large T might be necessary. In practice,

$$T = q + sQ + 20(p + sQ) \qquad [A12.2.6]$$

is often sufficient. The other truncation error involves terminating the iterative procedure used to calculate $[a_t]$. Several iterations may be required to obtain convergence when the model contains a MA factor with roots near the unit circle. However, sufficient accuracy is usually obtained on the first evaluation.

McLeod (1977) suggests that the term $M_n(p,q,P,Q,s)$ be replaced by $m(p,q,P,Q,s)$, given by

$$m(p,q,P,Q,s) = M(p,q)[M(P,Q)]^s, \qquad [A12.2.7]$$

where $M(p,q)$ is defined for any ARMA(p,q) model as

$$M(p,q) = M_p^2 M_q^2 / M_{p+q} \qquad [A12.2.8]$$

where the terms M_p, M_q and M_{p+q} are defined in terms of the auxiliary autoregressions, $\phi(B)v_t = a_t$ and $\theta(B)u_t = a_t$ and the left-adjoint autoregression $\phi(B)\theta(B)y_t = a_t$. For the autoregression, $\phi(B)v_t = a_t$, M_p is the determinant of the $p \times p$ matrix with (i,j) entry

$$\sum_{k=1}^{\min(i,j)} \phi_{i-k}\phi_{j-k} - \phi_{p+k-i}\phi_{p+k-j} \qquad [A12.2.9]$$

and similarly for the other autoregressions. The $p \times p$ matrix defined by [A12.2.9] is called the Schur matrix of $\phi(B)$. Pagano (1973) has shown that a necessary and sufficient condition for stationarity of an autoregression is that its Schur matrix be positive-definite (see Section 3.2.2). Thus, calculation of $m(p,q,P,Q,s)$ also provides a check on the stationarity and invertibility conditions and so during estimation the parameters may be constrained to the admissible region. Modified Cholesky decomposition is used to evaluate $M(p,q)$.

To obtain MLE's for the model parameters, the modified sum of squares must be minimized by using a standard optimization algorithm (see Section 6.2.3). McLeod and Salas (1983) describe in detail how the backforecasting method of Box and Jenkins (1976, Ch. 7) for ARMA models can be efficiently adapted for employment with SARMA models by making use of their multiplicative structure.

PROBLEMS

12.1 Select an average monthly riverflow time series that has at least ten years of data. Employ a suitable set of time series programs such as the McLeod-Hipel Package referred to in Section 1.7 to help you perform the following tasks.

(a) Examine a graph of the observations plotted over time and comment upon the overall statistical characteristics of the data. If necessary use other exploratory data analysis tools (see Sections 1.2.4, 5.3.2, 12.3.2 and 22.3) to study the statistical properties of your series.

(b) Based upon the discussion given at the start of Part VI and elsewhere in the chapter, which type of seasonal model do you feel is most appropriate to fit to the time series?

(c) Utilizing the techniques of Section 12.3, follow the three stages of model construction to develop the most appropriate SARIMA model to fit to this series. Show and explain all of your modelling results at each model building stage.

12.2 Carry out question 12.1 for a seasonal socio-economic time series of your choice. For example, you may wish to study a quarterly water or electrical demand series.

12.3 Execute question 12.1 for a seasonal water quality time series that has no missing values.

12.4 Select a seasonal meteorological time series for answering the questions given in question 12.1.

12.5 Write a SARIMA$(2,1,3)\times(1,1,2)_4$ model that is fitted to a series z_t with $\lambda = 0.5$ in the following forms:

(a) difference equation format given in [12.2.7],

(b) unfactored form in [12.2.10],

(c) generalized style as in [12.2.12],

(d) random shock form given in [12.2.15],

(e) and inverted format written in [12.2.17].

12.6 Derive the theoretical ACF for the SARMA$(p,q)\times(P,Q)$ model in [12.2.9].

12.7 For a SARIMA$(1,1,2)\times(1,1,1)_{12}$ model fitted to a series with $\lambda = 0.5$, calculate the MMSE forecasts for $l = 1,2, \ldots, 24$.

12.8 For the same SARIMA model given in 12.7, describe the steps for simulating 10,000 sequences of length 120 for this model. Where necessary, use equations to explain how calculations are carried out.

12.9 Select one of the calibrated SARIMA models that you fitted to a seasonal time series in problems 12.1 to 12.4. Calculate and plot the MMSE forecasts as well as the 90% probability intervals using this model for lead times from 1 to 25 where s is the seasonal length. Comment upon your results.

12.10 Choose one of the calibrated SARIMA models that you fitted to a seasonal data set in one of the first four questions. Simulate and plot three synthetic sequences of length $8s$ for the SARIMA model where s is the seasonal length. Compare these simulated sequences to the original series and discuss your findings.

12.11 Outline the procedure of Box et al. (1987) for estimating the trend in a seasonal time series that can be described using a SARIMA model. Comment upon the advantages and drawbacks of their approach.

12.12 Briefly describe how the tests of Li (1991) work for determining the orders of differencing required for modelling a seasonal time series. Comment upon the usefulness of the differencing tests.

REFERENCES

CARBON DIOXIDE DATA

Bacastow, R. B. and Keeling, C. D. (1981). Atmospheric carbon dioxide concentration and the observed airborne fraction. In Bolin, B., editor, *Carbon Cycle Modelling*, pages 103-112. John Wiley, Chichester.

Blue, C. (1991). CO_2 in glacial ice gives clues to historical atmospheres. *EOS, Transactions of the American Geophysical Union*, 72(36):379.

Keeling, C. D., Bacastow, R. B. and Whorf, T. P. (1982). Measurements of the concentration of carbon dioxide at Mauna Loa Observatory, Hawaii. In Clark, W. C., Editor, *Carbon Dioxide Review 1982*, pages 377-385. Clarendon Press, Oxford.

Neftel, A., Oeschger, H., Schwander, J., Stauffer, B. and Zumbrunn, R. (1982). Ice core sample measurements give atmospheric CO_2 content during the past 40,000 years. *Nature*, 295:216-219.

Pearman, G. I., Etheridge, D., de Silva, F. and Fraser, P. J. (1986). Evidence of changing concentrations of atmospheric CO_2, N_2O, and CH_4 from bubbles in Antarctic ice. *Nature*, 320:248-250.

Wahlen, M., Allen, D., Deck, B. and Herchenroder, A. (1991). Initial measurements of CO_2 concentrations (1530 to 1940 AD) in air occluded in the GISP2 ice core from Central Greenland. *Geophysical Research Letters*, 18(8):1457-1460.

RIVERFLOW AND WATER CONSUMPTION DATA

Environment Canada (1977). Historical streamflow summary, Ontario. Technical report, Inland Waters Directorate, Water Resources Branch, Ottawa, Canada.

Public Utilities Commission (1989). Water useage data for the city of London, Ontario. Technical report, Public Utilities Commission, P.O. Box 2700, London, Ontario.

SARIMA MODELLING

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716-723.

Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society*, Series B, 8:27-41.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26:211-252.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Box, G. E. P. and Pierce, D. A. (1970). Distribution of the residual autocorrelations in autoregressive integrated moving average models. *Journal of the American Statistical Association*, 65:1509-1526.

Box, G. E. P., Pierce, D. A. and Newbold, P. (1987). Estimating trend and growth rates in seasonal time series, *Journal of the American Statistical Association*, 82(397):276-282.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics*, 14(2):277-298.

Godolphin, E. J. (1977). On the autocorrelation structure for seasonal moving average models and its implications for the Cramer-Wold decomposition. *Journal of Applied Probability*, 14:785-794.

Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society*, Series B, 38(2):189-203.

Hamilton, D. C. and Watts, D. G. (1978). Interpreting partial autocorrelation functions of seasonal time series models. *Biometrika*, 65(1):135-140.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research*, 13(3):567-579.

Irvine, K. N. and Eberhardt, A. J. (1992). Multiplicative seasonal ARIMA models for Lake Erie and Lake Ontario water levels, *Water Resources Bulletin*, 28(2):385-396.

Li, W. K. (1991). Some Lagrange multiplier tests for seasonal differencing, *Biometrika*, 78(2):381-387.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, Series B, 43(2):231-239.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. *Biometrika*, 64(3):531-534.

McLeod, A. I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society*, Series B, 40(3):296-302.

McLeod, A. I. (1984). Duality and other properties of multiplicative seasonal autoregressive-moving average models. *Biometrika*, 71:207-211.

McLeod, A. I. (1993). Parsimony, model adequacy and periodic correlation in time series forecasting. *The International Statistical Review*, 61(3).

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, Applications. *Water Resources Research,* 13(3):577-586.

McLeod, A. I. and Salas, P. R. H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 32:211-223.

Ozaki, T. (1977). On the order determination of ARIMA models. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 26(3):290-301.

Pagano, M. (1973). When is an autoregressive scheme stationary? *Communications in Statistics,* 1(6):533-544.

CHAPTER 13

DESEASONALIZED MODELS

13.1 INTRODUCTION

Deseasonalized models are useful for describing time series in which the mean and variance within each season are stationary across the years. An example of a time series exhibiting these properties is the average monthly flows of the Saugeen River at Walkerton, Ontario, Canada, plotted in Figure VI.1. As can be seen from the sinusoidal shape of this graph, the mean and perhaps also the variance for each month change from season to season. Nevertheless, if one examines, for instance, the July flows across all of the years, these flows fluctuate around an overall mean level for the month of July and the variance is constant over time for the month of July. The deseasonalized models of this chapter and also the periodic models of Chapter 14 are ideally designed for capturing this type of statistical behaviour.

In addition to the *flexible design* of the deseasonalized models presented in Section 13.2, there are other distinct advantages for employing these models in practical applications. As explained in Section 13.3, a deseasonalized model can be easily identified and fitted to a given time series. Firstly, the seasonal component is removed from the series by subtracting from each observation the seasonal mean and perhaps also dividing this by the seasonal standard deviation. Subsequently, the most appropriate nonseasonal ARMA model is identified for fitting to the resulting nonseasonal series. Hence, the *model construction* tools presented in Part III for nonseasonal ARMA models, can also be used for building deseasonalized models.

To demonstrate clearly that the modelling methods can be easily used in practice, deseasonalized models are fitted to two *environmental time series* in Section 13.4. In the first application, a deseasonalized model is built for the average monthly flows of the Saugeen River displayed in Figure VI.1 while in the second example the most appropriate deseasonalized model is constructed for an average monthly ozone series.

Two other advantages for employing deseasonalized models are that they can be easily used for forecasting and simulation. As explained in Section 13.5, when *forecasting* with a deseasonalized model, the nonseasonal component of the series is forecasted using the procedure of Chapter 8 for nonseasonal ARMA models. Next, these forecasted values are converted to seasonal forecasts using the inverse of the deseasonalization procedure. Finally, if the original data were also transformed using a Box-Cox transformation, the inverse Box-Cox transformation is invoked to produce forecasts which possess the same units as the original untransformed series. A similar procedure is employed for *simulating* synthetic data using a deseasonalized model. The techniques of Chapter 9 are used to generate the simulated data from the nonseasonal ARMA model fitted to the deseasonalized series. Following this, the nonseasonal simulated values are filtered through the inverse deseasonalization method and inverse Box-Cox transformation to produce the untransformed synthetic sequence.

A drawback of using deseasonalized models is that they may require many model parameters for deseasonalizing the data. For instance, when dealing with a monthly time series, twelve monthly means and perhaps also twelve monthly standard deviations are needed for deseasonalization purposes. When modelling bimonthly and weekly time series a total of 48 and 104 deseasonalization parameters, respectively, may be required. To overcome the problem of *over-parameterization*, a *Fourier series approach* is described in Section 13.2.2. The applications in Section 13.4 are used for explaining how this method for reducing the number of model parameters is used in practice by following the overall model building stages explained in Section 13.3. In Section 13.5, approaches for forecasting and simulating with deseasonalized models are explained.

13.2 DEFINITIONS OF DESEASONALIZED MODELS

13.2.1 Introduction

As noted in the introduction, the deseasonalized model consists of the deseasonalization and ARMA model components. Because the basic design of the deseasonalized model reflects the inherent structure of many kinds of seasonal hydrological time series, this model has been used by hydrologists and environmental engineers for a long time for modelling, simulating and forecasting hydrologic phenomena (see, for example, Thomas and Fiering (1962), McMichael and Hunter (1972), McKerchar and Delleur (1974), Kavvas and Delleur (1975), Delleur et al. (1976), Tao and Delleur (1976), Croley and Rao (1977), Yevjevich and Harmancioglu (1989), and Jaywardena and Lai (1989), as well as the books on stochastic hydrology referred to in Section 1.6.3). In addition to defining this popular type of seasonal model in Sections 13.2.2 and 13.2.3, the AIC (Akaike information criterion) formula for the deseasonalized model is determined in Section 13.3.3 following the research of Hipel and McLeod (1979).

13.2.2 Deseasonalization

When considering data with s seasons per year ($s = 12$ for monthly data) over a period of n years, let $z_{r,m}$ represent a time series value in the rth year and mth season where $r = 1,2,\ldots,n$, and $m = 1,2,\ldots,s$. It is convenient to denote the ith previous value of $z_{r,m}$ by $z_{r,m-i}$, $i = 1,2,\ldots$. If, for example, one were dealing with monthly data, then $z_{9,12}$, $z_{10,0}$, and $z_{8,24}$ would all refer to the same observation.

If required, the given data may be transformed by the *Box-Cox transformation* (Box and Cox, 1964) to form the transformed series

$$z_{r,m}^{(\lambda)} = \begin{cases} \lambda^{-1}[(z_{r,m}+c)^{\lambda}-1], & \lambda \neq 0 \\ \ln(z_{r,m}+c), & \lambda = 0 \end{cases} \qquad [13.2.1]$$

for $r = 1,2,\ldots,n$, and $m = 1,2,\ldots,s$, where the constant c is chosen just large enough to cause all entries in $z_{r,m}^{(\lambda)}$ to be positive, and λ is the Box-Cox power transformation. Although one could have a separate Box-Cox transformation for each season of the year, in order to reduce the number of parameters it is assumed that the same λ and c are used for each season. The purpose of the Box-Cox transformation is to rectify anomalies such as heteroscedasticity and non-normality in the residuals of the ARMA model fitted to the deseasonalized time series.

The Box-Cox transformation given in [13.2.1] cannot remove seasonality from a time series. However, the following equations describe two *deseasonalization methods* that can be employed to deseasonalize the given data.

$$w_{r,m}^{(1)} = z_{r,m}^{(\lambda)} - \bar{\mu}_m \qquad [13.2.2]$$

$$w_{r,m}^{(2)} = \frac{z_{r,m}^{(\lambda)} - \bar{\mu}_m}{\bar{\sigma}_m} \qquad [13.2.3]$$

where $\bar{\mu}_m$ and $\bar{\sigma}_m$ are the fitted mean and standard deviation for the mth season.

The deseasonalization procedures given in [13.2.2] and [13.2.3], reflect the inherent statistical realities of many kinds of natural time series. For example, when considering average monthly river flow data, the observations for any particular month tend to fluctuate about some fixed mean level. Consequently, the deseasonalization method in [13.2.2] may be appropriate to employ if the monthly standard deviations of the $z_{r,m}^{(\lambda)}$ series are more or less constant throughout the year. When both the means and standard deviations of the $z_{r,m}^{(\lambda)}$ sequence are different from month to month, then the transformation in [13.2.3] should be utilized. In certain situations, a Box-Cox transformation may cause the standard deviations to become constant throughout the year for the $z_{r,m}^{(\lambda)}$ series and hence [13.2.2] can be used in preference to [13.2.3]. However, as pointed out in Chapter 12, for the type of natural time series just described, it is not recommended to difference the data to remove seasonality if the model is to be used for simulation.

The fitted means and standard deviations in [13.2.2] and [13.2.3] can be estimated using two approaches. One method is to estimate them using the standard formulae. Hence, the estimate, $\hat{\mu}_m$, for the mean of the mth season across all of the years, can be determined using

$$\hat{\mu}_m = \frac{1}{n}\sum_{r=1}^{n} z_{r,m}^{(\lambda)}, \quad m = 1,2,\ldots,s \qquad [13.2.4]$$

The estimate of the standard deviation for the mth season, is calculated as

$$\hat{\sigma}_m = \left[\frac{1}{n}\sum_{r=1}^{n}(z_{r,m}^{(\lambda)} - \hat{\mu}_m)^2\right]^{0.5}, \quad m = 1,2,\ldots,s \qquad [13.2.5]$$

For the case of monthly data, the deseasonalization transformation in [13.2.2] would require 12 means whereas the one in [13.2.3] needs 12 means and 12 standard deviations for a total of 24 deseasonalization parameters. When dealing with bimonthly, weekly and daily observations, a far greater number of deseasonalization parameters would be required

To reduce the total number of deseasonalization parameters that are needed, a ***Fourier series approach to deseasonalization*** can be utilized as another method for estimating the means and standard deviations. Let F_μ and F_σ be the number of Fourier components to fit to the seasonal means in [13.2.4] and seasonal standard deviations in [13.2.5], respectively. Both F_μ and F_σ can possess an integer value between 0 and $s/2$. The Fourier coefficients are determined from the equations.

$$A_k = \frac{2}{s}\sum_{m=1}^{s}\hat{\mu}_m\cos\frac{2\pi km}{s} \qquad [13.2.6]$$

$$B_k = \frac{2}{s}\sum_{m=1}^{s}\hat{\mu}_m\sin\frac{2\pi km}{s} \qquad [13.2.7]$$

$$C_h = \frac{2}{s}\sum_{m=1}^{s}\hat{\sigma}_m\cos\frac{2\pi hm}{s} \qquad [13.2.8]$$

$$D_h = \frac{2}{s}\sum_{m=1}^{s}\hat{\sigma}_m\sin\frac{2\pi hm}{s} \qquad [13.2.9]$$

where $k = 1,2, \ldots, F_\mu$, and $h = 1,2, \ldots, F_\sigma$ for $1 \le F_\mu$, $F_\sigma \le s/2$. To calculate the fitted means and standard deviations, set $\bar{\mu}_m = 0$ if $F_\mu = 0$ and let $\bar{\sigma}_m = 1$ if $F_\sigma = 0$. Otherwise, determine $\bar{\mu}_m$ and $\bar{\sigma}_m$ using

$$\bar{\mu}_m = A_0 + \sum_{k=1}^{F_\mu}\left[A_k\cos\frac{2\pi km}{s} + B_k\sin\frac{2\pi km}{s}\right], \quad m = 1,2, \ldots, s \qquad [13.2.10]$$

$$\bar{\sigma}_m = C_0 + \sum_{h=1}^{F_\sigma}\left[C_h\cos\frac{2\pi hm}{s} + D_h\sin\frac{2\pi hm}{s}\right], \quad m = 1,2, \ldots, s \qquad [13.2.11]$$

where $A_0 = \frac{1}{s}\sum_{m=1}^{s}\hat{\mu}_m$, and $C_0 = \frac{1}{s}\sum_{m=1}^{s}\hat{\sigma}_m$.

The deseasonalized series can be calculated using [13.2.2] if $F_\sigma = 0$ or otherwise [13.2.3]. When all of the Fourier components are used to calculate $\bar{\mu}_m$ (or $\bar{\sigma}_m$), then $\bar{\mu}_m = \hat{\mu}_m$ (and $\bar{\sigma}_m = \hat{\sigma}_m$) and, hence, $\bar{\mu}_m$ in [13.2.10] is equal to $\hat{\mu}_m$ in [13.2.4] and $\bar{\sigma}_m$ in [13.2.11] is the same as $\hat{\sigma}_m$ in [13.2.5]. Therefore, estimating the means and standard deviations using [13.2.4] and [13.2.5], respectively, can be considered a special case of the Fourier series approach.

13.2.3 ARMA Model Component

Subsequent to deseasonalization, the ARMA model defined in [3.4.4] can be fitted to the deseasonalized series represented by $w_{r,m}^{(1)}$ or $w_{r,m}^{(2)}$ in [13.2.2] or [13.2.3], respectively. For the case of the $w_{r,m}^{(2)}$ series in [13.2.3], the ARMA model would be written as

$$\phi(B)w_{r,m}^{(2)} = \theta(B)a_{r,m} \qquad [13.2.12]$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B - \cdots - \phi_p B^p$ is the AR operator of order p for which ϕ_i is the ith AR parameter and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the MA operator of order q for which θ_j is the jth MA parameter. The innovation series, $a_{r,m}$, is assumed to be distributed as IID$(0,\sigma_a^2)$. For estimating the model parameters, the normality assumption is invoked so that the innovations are required to be NID$(0,\sigma_a^2)$. If the residuals of the fitted model are not normally distributed and/or do not possess a constant variance, then an appropriate power transformation from [13.2.1] should be selected for correcting these problems.

13.3 CONSTRUCTING DESEASONALIZED MODELS

13.3.1 Introduction

When constructing a deseasonalized model, one follows in a general fashion the identification, estimation and diagnostic check stages of model building. During model construction, one must be able to identify and estimate the parameters contained in the deseasonalization and ARMA model components which form the overall deseasonalized model. In the next two sections, two basic approaches are presented for carrying out the model building process. In the first method, the time series under study is fully deseasonalized using [13.3.4] and [13.3.5] to estimate the deseasonalization parameters given in [13.2.2] and [13.2.3]. Then the best ARMA model is fitted to the resulting nonseasonal time series. The resulting model is then subjected to diagnostic check to ascertain if design modifications are needed to correct abnormalities in the ARMA model residuals. In the second approach, the AIC is used to determine which Fourier parameters are required in the deseasonalization step so that a more parsimonious model can be found. This second method is often needed when one is examining series, such as bimonthly or weekly sequences, for which the number of seasons per year is greater than twelve. Special AIC formulae are derived for deseasonalized models so one can employ the MAICE (minimum AIC estimation) procedure (see Section 6.3) for selecting the model having the fewest number of parameters. The AIC formulae are also valid for use with deseasonalized models having a Box-Cox transformation.

Figure 13.3.1 outlines the two procedures for fitting deseasonalized models to a time series. When following the fully deseasonalized approach of Section 13.3.2, one does not use the AIC to determine the Fourier components needed in the deseasonalization. Alternatively, when utilizing the AIC to determine which Fourier components are needed, one is constructing the deseasonalized model according to the procedure explained in Section 13.3.3. Notice that Figure 13.3.1 is similar to the general AIC model building methodology depicted in Figure 6.3.1. Moreover, most of the model building tools of Part III, can be employed for developing the ARMA component of the deseasonalized model of Sections 13.3.2 and 13.3.3.

13.3.2 Fully Deseasonalized Models

As just mentioned, the approach for fitting a *fully deseasonalized model* to a seasonal time series is found in Figure 13.3.1 by following the path which ignores using the AIC for determining the Fourier components needed in the deseasonalization method. Firstly, if it is known a priori that a Box-Cox transformation is needed, one can use [13.2.1] to transform the data. If the need for a Box-Cox transformation is not known in advance, it will be detected by testing the model residuals at the diagnostic check stage. Subsequent to implementing an appropriate transformation, the model parameters can be calibrated once again as shown in Figure 13.3.1.

Secondly, the transformed seasonal series is fully deseasonalized. If it is suspected based on graphs of the series or estimates of the seasonal standard deviations that the seasonal variance is about the same across the seasons, one can use [13.2.2] to deseasonalize the series. When using [13.2.2], each seasonal mean is estimated using the standard formula given in [13.2.4]. If, on the other hand, the variance changes across the seasons, [13.2.3] is used for obtaining the deseasonalized series where [13.2.4] and [13.2.5] are utilized for estimating the seasonal means and standard deviations, respectively.

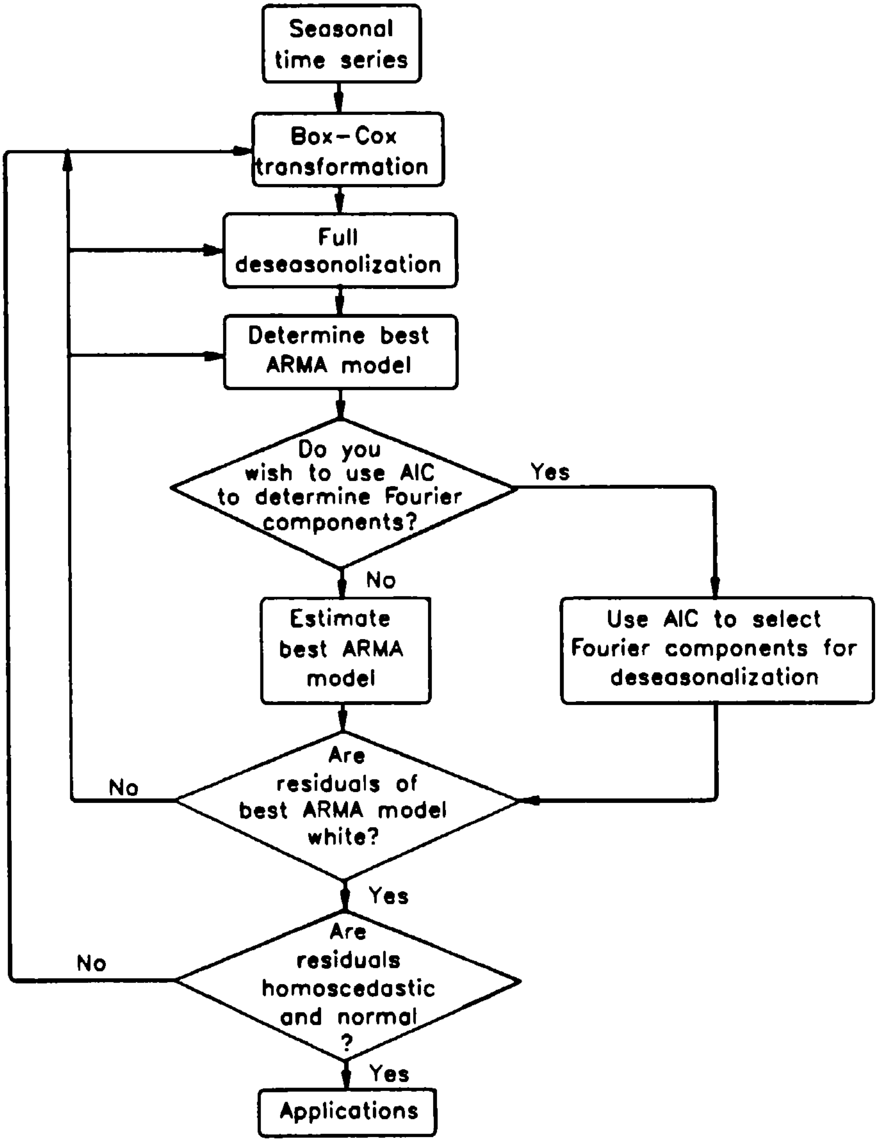

Seasonal time series
Box-Cox transformation
Full deseasonalization
Determine best ARMA model
Do you wish to use AIC to determine Fourier components?
Yes
No
Estimate best ARMA model
Use AIC to select Fourier components for deseasonalization
Are residuals of best ARMA model white?
No
Yes
Are residuals homoscedastic and normal ?
No
Yes
Applications

Figure 13.3.1. Constructing deseasonalized models.

Thirdly, the procedures of Part III are used to determine the best ARMA model to fit to the nonseasonal series found using [13.2.2] or [13.2.3]. If, for example, the best ARMA is not clearly identified using the identification graphs presented in Chapter 5, one may have to estimate the parameters for a few different ARMA models and select the one possessing the minimum value of the AIC (see Section 6.3). Whatever the case, in Figure 13.3.1, one omits using the AIC for determining the Fourier components and proceeds to estimating the parameters of the most appropriate ARMA model. Next, the residuals of the fitted ARMA model are checked to see if their underlying assumptions are satisfied. In particular, one should make sure that the residuals are uncorrelated, normally distributed and homoscedastic. Specific tests for checking that these properties are not violated are presented in detail in Chapter 7. Problems with heteroscedasticity and/or non-normality can be corrected by invoking a suitable Box-Cox transformation. If the residuals are correlated, a different ARMA model should be fitted to the fully deseasonalized series. After obtaining a satisfactory model, the overall deseasonalized model can be used for application purposes such as forecasting and simulation.

13.3.3 Fourier Approach to Deseasonalized Models

Overall Procedure

The overall procedure for carrying out the *Fourier approach to deseasonalized models* is explained first. Following this, the AIC formulae which are used in this procedure are derived.

The Fourier approach is traced in Figure 13.3.1 by following the path which uses the AIC for ascertaining the Fourier components needed in the deseasonalization. As can be seen, the first three steps consisting of a Box-Cox transformation, full deseasonalization, and determining the best ARMA model to fit to the fully deseasonalized data are identical to those explained for the fully deseasonalized model in Section 13.3.2. To cut down on the number of deseasonalization parameters that are needed in either [13.2.2] or [13.2.3], the MAICE procedure can be used to select the Fourier components that are required in [13.2.10] to [13.2.11]. When doing this, one can assume that the ARMA model identified for fitting to the fully deseasonalized data set requires the same number of AR and MA parameters. For each possible combination of Fourier components, one can estimate the AIC for a given type of ARMA model. The deseasonalized model possessing the minimum AIC value is then selected as the most satisfactory model. If one suspected that the type of ARMA model fitted to the deseasonalized data were dependent upon the number of Fourier components, one could also allow the number of AR and MA parameters to vary during the MAICE procedure. However, this would require a significant increase in the amount of computations and therefore is not shown in Figure 13.3.1. Whatever the case, subsequent to choosing the best deseasonalized model, one can subject the ARMA model residuals to the diagnostic checks given in Chapter 7. As shown in Figure 13.3.1, problems with the model residuals can be rectified by returning to an earlier step in the modelling procedure. The best overall deseasonalized model can then be used for application purposes.

AIC Formulae for Deseasonalized Models

Recall from [6.3.1] Section 6.3, that the general formula for the AIC is given as (Akaike, 1974)

$$AIC = -2\ln(ML) + 2k$$

where ML stands for the maximized value of the likelihood function and k is the number of free parameters. The model which adequately fits a time series using a minimum number of parameters possesses the lowest AIC value. The procedure for determining the model having the minimum AIC value is referred to as MAICE.

When calculating the AIC formula for the deseasonalized model, one must consider the effects of a Box-Cox transformation and deseasonalization. Hence, one must determine the Jacobians of the deseasonalization transformations.

Following the derivation given by Hipel and McLeod (1979), first consider the deseasonalization procedure given in [13.2.2]. The Jacobian of the transformation from $z_{r,m}$ to $w_{r,m}^{(1)}$ is

$$J_1 = \left| \left(\frac{\partial w_{r,m}^{(1)}}{\partial z_{r,m}} \right) \right| = \prod_{r=1}^{n} \prod_{m=1}^{s} z_{r,m}^{\lambda-1} \qquad [13.3.1]$$

To explain in detail how [13.3.1] is derived, examine the Box-Cox transformation in [13.2.1]. Assuming that all of the $z_{r,m}$ are positive and, therefore, the constant is zero,

$$z_{r,m}^{(\lambda)} = \frac{z_{r,m}^{\lambda} - 1}{\lambda}$$

and from [13.2.2] the deseasonalized transformed series is

$$w_{r,m}^{(1)} = \frac{z_{r,m}^{\lambda} - 1}{\lambda} - \bar{\mu}_m$$

The entire matrix used in the Jacobian calculation in [13.3.1] is

$$\begin{pmatrix} \frac{\partial w_{11}^{(1)}}{\partial z_{11}} & \frac{\partial w_{11}^{(1)}}{\partial z_{12}} & \cdots & \frac{\partial w_{11}^{(1)}}{\partial z_{n,s}} \\ \frac{\partial w_{12}^{(1)}}{\partial z_{11}} & \frac{\partial w_{12}^{(1)}}{\partial z_{12}} & \cdots & \frac{\partial w_{12}^{(1)}}{\partial z_{n,s}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial w_{r,s}^{(1)}}{\partial z_{11}} & \frac{\partial w_{r,s}^{(1)}}{\partial z_{12}} & \cdots & \frac{\partial w_{r,s}^{(1)}}{\partial z_{n,s}} \end{pmatrix}$$

In this matrix, all of the off-diagonal entries have a value of zero whereas each diagonal element is evaluated as

$$\frac{\partial w_{r,m}^{(1)}}{\partial z_{r,m}} = \frac{\lambda z_{r,m}^{\lambda-1}}{\lambda} = z_{r,m}^{\lambda-1}$$

Therefore, the value of the determinant of this matrix which constitutes the Jacobian J_1 is found by multiplying the diagonal entries together to obtain the result in [13.3.1].

The natural logarithm of J_1 is

$$\ln(J_1) = (\lambda - 1)\sum_{r=1}^{n}\sum_{m=1}^{s}\ln(z_{r,m}) \qquad [13.3.2]$$

The log-likelihood in terms of $z_{r,m}$ is then

$$\ln(L^{(1)}) \approx -ns\ln\left[\frac{MSS}{ns}\right] + \ln(J_1) \qquad [13.3.3]$$

where MSS stands for the modified sum of squares described in Appendix A6.1.

The AIC for an ARMA model that is fit to the $w_{i,j}^{(1)}$ series is

$$AIC = -2\ln(L^{(1)}) + 2(p + q + 2 + \delta_1) + 4F_\mu - 2\delta_2 \qquad [13.3.4]$$

where p is the number of AR parameters, q is the number of MA parameters, $\delta_1 = 0$ when $\lambda = 1$ and $\delta_1 = 1$ when $\lambda \neq 1$, F_μ is the number of Fourier parameters in [13.2.10] used to estimate the seasonal means, and $\delta_2 = 1$ if $F_\mu = \frac{s}{2}$ while $\delta_2 = 0$ when $F_\mu < \frac{s}{2}$.

In the deseasonalization method given in [13.2.3], each entry in the $w_{r,m}^{(1)}$ series is divided by the appropriate standard deviation of $z_{r,m}^{(\lambda)}$ for the mth season. The Jacobian of the transformation from $w_{r,m}^{(1)}$ to $w_{r,m}^{(2)}$ is

$$J_2 = \left|\left(\frac{\partial w_{r,m}^{(2)}}{\partial w_{r,m}^{(1)}}\right)\right| = \left(\prod_{m=1}^{s}\bar{\sigma}_m^{-1}\right)^n \qquad [13.3.5]$$

To derive [13.3.5], consider the matrix used in the Jacobian formulae in [13.3.5] which is written as

$$\begin{bmatrix} \frac{\partial w_{11}^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial w_{11}^{(2)}}{\partial w_{12}^{(1)}} & \cdots & \frac{\partial w_{11}^{(2)}}{\partial w_{n,s}^{(1)}} \\ \frac{\partial w_{12}^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial w_{12}^{(2)}}{\partial w_{12}^{(1)}} & \cdots & \frac{\partial w_{12}^{(2)}}{\partial w_{n,s}^{(1)}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial w_{n,s}^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial w_{n,s}^{(2)}}{\partial w_{12}^{(1)}} & \cdots & \frac{\partial w_{n,s}^{(2)}}{\partial w_{n,s}^{(1)}} \end{bmatrix}$$

From [13.2.2] and [13.2.3], the relationship between $w_{r,m}^{(1)}$ and $w_{r,m}^{(2)}$ determined as

$$w^{(2)}_{r,m} = \frac{W^{(1)}_{r,m}}{\hat{\sigma}_m}$$

Hence,

$$\frac{\partial w^{(2)}_{r,m}}{\partial w^{(1)}_{r,m}} = \frac{1}{\hat{\sigma}_m} = \hat{\sigma}_m^{-1}$$

Therefore, the diagonal elements in the matrix have a value of $\hat{\sigma}_m^{-1}$ while the off-diagonal elements are zero. The determinant of this matrix which is used to obtain the Jacobian J_2 is determined simply by multiplying together the values of the diagonal elements to obtain [13.3.5].

The natural logarithm of J_2 is

$$\ln(J_2) = -n \sum_{m=1}^{12} \ln(\hat{\sigma}_m) \qquad [13.3.6]$$

The log-likelihood for the $z_{r,m}$ series is

$$\ln(L^{(2)}) \approx -ns\ln\left[\frac{MSS}{ns}\right] + \ln(J_1) + \ln(J_2) \qquad [13.3.7]$$

The AIC formula for an ARMA model that is fitted to the $w^{(2)}_{r,m}$ sequence is

$$AIC = -2\ln(L^{(2)}) + 2(p + q + 2 + \delta_1) + 4(F_\mu + F_\sigma) - 2(\delta_2 + \delta_3) \qquad [13.3.8]$$

where F_σ is the number of Fourier parameters in [13.2.11] used to estimate the seasonal standard deviations, $\delta_3 = 1$ when $F_\sigma = \frac{s}{2}$ while $\delta_3 = 0$ whenever $F_\sigma < \frac{s}{2}$.

For certain data sets, it may be known in advance what type of Box-Cox transformation should be used. Hence, it may be appropriate to set λ at a specified value. In other situations, it may be desirable to obtain a maximum likelihood estimate for the Box-Cox exponent. This can be accomplished by maximizing the log-likelihood in [13.3.3] and [13.3.7] with respect to the model parameters.

As explained in Section 6.3.4, the difference in the values of the AIC for various models which are fit to the same data set, can be interpreted in different manners. For instance, if one model has an AIC value which is approximately $2k$ less than that for another model, this is analogous to the superior model having k less parameters than the other model. An alternative approach for interpreting the difference in AIC values between two models, is to determine the plausibility of model i versus model j by using the formula

$$\text{Plausibility} = \exp[0.5(AIC_j - AIC_i)] \qquad [13.3.9]$$

where AIC_i is the value of the AIC for the ith model, AIC_j is the value of the AIC for the jth model and the jth model is assumed to be the model having the lower AIC value. This formula was suggested by H. Akaike in a private communication and is also written in [6.3.2].

13.4 APPLICATIONS OF DESEASONALIZED MODELS

13.4.1 Introduction

To demonstrate how the model construction approaches of Section 13.3 are used in practice, deseasonalized models are developed for two natural time series by adhering to the procedures outlined in Figure 13.3.1. In the first application, a deseasonalized model is fitted to the average monthly flows of the Saugeen River shown in Figure VI.1. The MAICE procedure is utilized in the second example for determining a deseasonalized model to describe a monthly ozone time series. Ozone is used as an indicator of pollution levels caused by exhausts from automobiles and other types of machinery driven by internal combustion engines. By using the MAICE procedure, very few deseasonalization parameters are needed in the fitted deseasonalized model.

13.4.2 Average Monthly Saugeen Riverflows

The average monthly flows of the Saugeen River at Walkerton, Ontario, Canada, are available from Environment Canada (1977) from January, 1915, until December, 1976, and the last ten years of this time series are plotted in Figure VI.1. This particular data set is suitable for modelling using a deseasonalized model because there are no major reservoirs on the river and also the land-use activities in the river basin have not changed significantly during the aforesaid time period. Consequently, the technique of intervention analysis of Part VIII does not have to be used to account for intervention effects.

First consider using the modelling construction approach of Section 13.3.2. This full deseasonalization procedure follows the vertical path on the left in Figure 13.3.1 which does not use the AIC for determining the optimum number of Fourier components needed in deseasonalization.

As pointed out in Section 12.4.4, in practice, it has been found necessary to first take natural logarithms of average monthly riverflow time series values. The logarithmic transformation often precludes problems with non-normality and/or heteroscedasticity in the residuals of the model which is fitted to the deseasonalized data. A common procedure in water resources engineering is to fully deseasonalize the transformed data by using the estimated means and standard deviations from [13.2.4] and [13.2.5], respectively, in place of the fitted means and standard deviations in [13.2.3]. This is equivalent to setting $F_{\mu} = F_{\sigma} = 6$ in [13.2.6] to [13.2.9]. Even though this may not constitute the most appropriate deseasonalization method for many time series, it is, however, useful for the model identification suggested in Figure 13.3.1. In addition, as shown by this example, most of the Fourier components are needed for deseasonalization. After taking natural logarithms of the monthly Saugeen River data which is given in cubic metres per second, the logarithmic data is deseasonalized following [13.2.3] by using the estimated means and standard deviations in [13.2.4] and [13.2.5], respectively. By following the identification procedures described in Chapter 5 for nonseasonal ARMA models, it is found that a model with one AR and one MA parameter [denoted as ARMA(1,1)] may adequately model the deseasonalized series.

Figure 13.4.1 displays the fully deseasonalized series for the last ten years of the average monthly riverflows of the Saugeen River. In order to be able to compare this figure to original Saugeen flows in Figure VI.1, the deseasonalized series in Figure 13.4.1 is determined for the situation where there is no Box-Cox transformation and hence $\lambda = 1.0$ in [13.2.1] as well as the deseasonalization calculation in [13.2.3]. As can be seen, the deseasonalized series in Figure 13.4.1 does not contain the distinct sinusoidal patterns reflecting seasonality shown in Figure VI.1. Finally, as noted above, the reader should keep in mind that the ARMA(1,1) model is identified for the fully deseasonalized series with $\lambda = 0$.

Table 13.4.1. AIC values for the ARMA(1,1) model fitted to the deseasonalized Saugeen River series.

F_μ	F_σ	AIC
5	0	3383.15
5	1	3359.17
5	2	3365.98
5	3	3357.93
5	4	3355.82
5	5	3356.77
5	6	3361.81
6	0	3387.13
6	1	3363.15
6	2	3369.96
6	3	3361.91
6	4	3359.81
6	5	3360.75
6	6	3365.79

At the estimation stage of model development, the estimation procedure described in Appendix A6.1 is employed to obtain MLE's for the parameters of the ARMA(1,1) model which is fitted to the fully deseasonalized data in [13.2.3]. Using [13.3.8], the AIC for this fully deseasonalized model can be calculated for the case of full deseasonalization which is equivalent to $F_\mu = F_\sigma = 6$ in the Fourier series approach to deseasonalization. The value of the AIC for the fully deseasonalized model is given at the bottom of Table 13.4.1. This model passes the diagnostic checks suggested in Figure 13.3.1 and described in detail in Chapter 7. Consequently, the best ARMA model to fit to fully deseasonalized logarithmic monthly Saugeen flows is an ARMA(1,1) model.

Now consider how the Fourier approach to deseasonalization presented in Section 13.3.3 is used with the Saugeen data. As depicted in Figure 13.3.1, one must now fit the ARMA(1,1) model to each deseasonalized data set formed by all possible combination of Fourier components for estimating the seasonal means and standard deviations. Hence, the ARMA(1,1) model is fitted to each of the possible 49 deseasonalized data sets and [13.3.8] is used to calculate the AIC for each case. Table 13.4.1 lists the values of the AIC for some of the deseasonalized data sets. As can be seen from that table, the minimum value of the AIC occurs when 5 Fourier components are used for the means while the standard deviations require 4 Fourier components. Notice that the deseasonalization procedure which uses 6 Fourier components for both the means and standard deviations, is somewhat inferior to the best method. From [13.3.9], the plausibility

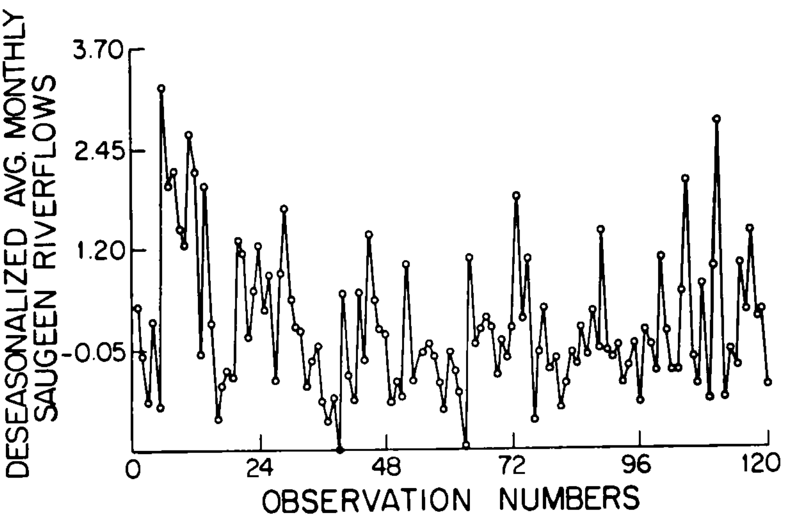


Figure 13.4.1. Fully deseasonalized series with λ = 1 in [13.2.3] for the last ten years of the average monthly flows of the Saugeen River at Walkerton, Ontario, Canada.

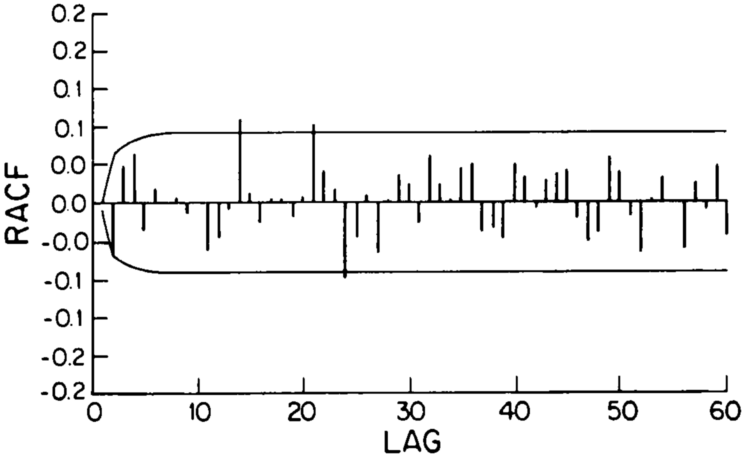


Figure 13.4.2. RACF for the ARMA(1,1) model fitted to the deseasonalized average monthly flows of the Saugeen River at Walkerton, Canada, with λ = 0, $F_\mu = 5$ and $F_\sigma = 4$.

of the model where the estimated means and standard deviations are used, versus the best seasonal model, is calculated to be 0.68%. Consequently, water resources engineers should not necessarily assume that the best deseasonalization method is to let $F_\mu = F_\sigma = 6$.

Examination of the residuals for the ARMA(1,1) model which is fitted to the deseasonalized data with $F_m = 5$ and $F_s = 4$, reveals that the whiteness assumption is satisfied. In particular, the graph of the RACF (residual autocorrelation function) in Figure 13.4.2 for this model shows that the values of the RACF fall within the 95% confidence limits. As discussed in Chapter 7 and elsewhere in this text, it has been found in practice that the important residual assumptions are usually fulfilled for the model which has been chosen using the MAICE procedure.

Diagnostic checks are given in Sections 7.4 and 7.5 to determine if the model residuals are approximately normally distributed and homoscedastic, respectively. If either of the aforesaid assumptions are not satisfied, the Box-Cox transformation in [13.2.1] can be invoked to rectify the situation. For the case of the Saugeen River data, it is assumed from the outset that a logarithmic transformation may be needed. Nevertheless, various values of λ in [13.2.1] are examined and $\lambda = 0$ does indeed constitute a reasonable transformation.

As would be expected all the deseasonalized models in Table 13.4.1 possess lower values than the AIC for the best SARIMA model fitted to the average monthly Saugeen Riverflows in Section 12.4.4. This is because the design of the deseasonalized model allows for a stationary mean and standard deviation within each season. The reader should keep in mind that because different estimation procedures are used for obtaining estimates of the model parameters for different classes of models, one should entertain caution when comparing AIC values across families of models.

13.4.3 Ozone Data

Monthly values of the concentration of ozone (in parts per hundred million) at Azusa, California, are available from January 1956 until December 1970. Figure 13.4.3 shows a plot of the ozone data, which are available in a technical report by Tiao et al. (1973). Abraham and Box (1978) have analyzed this data and have suggested that a moving average model with one parameter [denoted ARMA(0,1)] and no transformation (i.e., $\lambda = 1$ in [13.2.1]) is appropriate to model the deseasonalized data in [13.2.2] with $F_\mu = 2$. However, by employing the MAICE procedure in conjunction with the AIC formulae developed in Section 13.3.3, an improved ozone model is obtained.

Following Figure 13.3.1, a tentative model is identified to fit to the possible array of deseasonalized data sets by first examining the $w_{i,j}^{(2)}$ series in [13.2.3] with $F_\mu = F_\sigma = 6$. The fully deseasonalized series for the ozone data is displayed in Figure 13.4.3 and, as can be seen, deseasonalization removes the periodic seasonal component shown in Figure 13.4.3. A perusal of the sample ACF and PACF for this series, reveals that an AR model with one AR parameter [denoted ARMA(1,0)] may be suitable for modelling the data. The ARMA(1,0) model is fitted to the various types of deseasonalized time series which are obtained by simultaneously allowing F_μ and F_σ to take on all possible values between 0 and 6. The values of the AIC obtained using [13.3.8] are listed in Table 13.4.2 for some of the models considered. It can be seen that an ARMA(1,0) with no Box-Cox transformation (i.e. $\lambda = 1$) and $F_\mu = 2$ and $F_\sigma = 1$ is preferable to

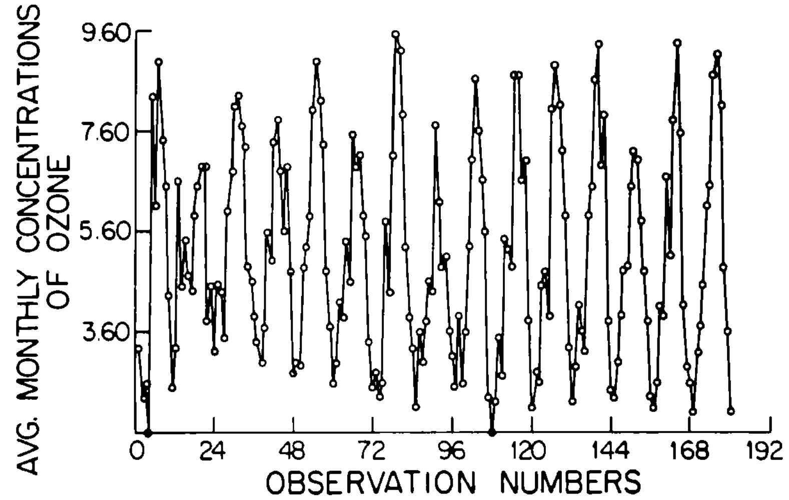


Figure 13.4.3. Average monthly concentrations of ozone (parts per hundred million) from January, 1956, to December, 1970, at Azusa, California.

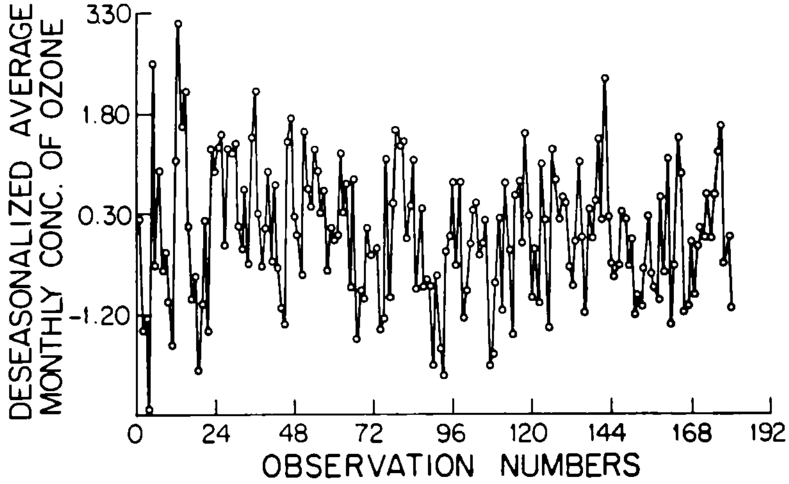


Figure 13.4.4. Fully deseasonalized series with $\lambda = 1$ in [13.2.3] for the last ten years of the average monthly concentrations of ozone (parts per hundred million) from January, 1956, to December, 1970, at Azusa, California.

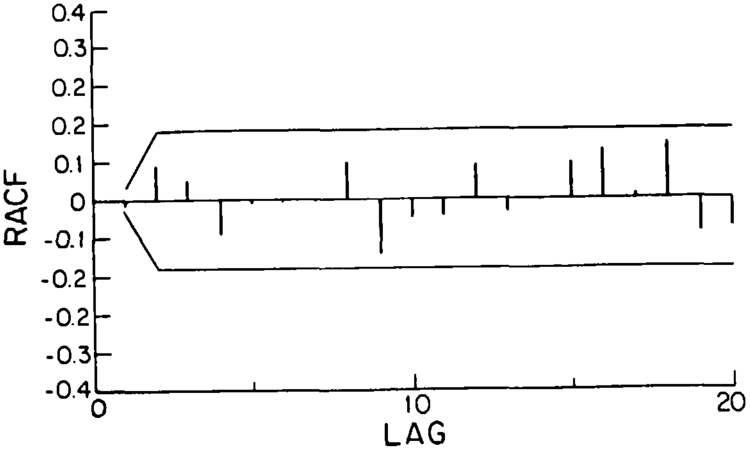


Figure 13.4.5. RACF for the ARMA(1,0) model fitted to the deseasonalized average monthly concentrations of ozone (parts per hundred million) at Azusa, California, with $\lambda = 0.5$, $F_{\mu} = 2$ and $F_{\sigma} = 1$.

the case where the same model is fitted to the deseasonalized data with $F_{\mu} = 2$ and $F_{\sigma} = 0$. In addition, as shown in Table 13.4.2, both of the aforesaid models possess lower AIC values than those obtained for ARMA(0,1) models which are fitted to the same deseasonalized series. However, an examination of the residuals for the most appropriate foregoing model, reveals that the residuals possess both significant skewness and heteroscedasticity. Consequently, various values of λ are examined in order to determine a suitable Box-Cox transformation from [13.2.1]. A transformation with $\lambda = 0.5$ corrects the anomolies in the model residuals. The graph of the RACF in Figure 13.4.5 for the ARMA(1,0) model fitted to the deseasonalized series with $\lambda = 0.5$, $F_{\mu} = 2$ and $F_{\sigma} = 1$ demonstrates that the residuals are white.

In Table 13.4.2, the values of the AIC are listed for various models which are fitted to the data having a square root transformation. It is evident that the most desirable model is an ARMA(1,0) model with $\lambda = 0.5$, $F_{\mu} = 2$ and $F_{\sigma} = 0$. The plausibility of the process recommended by Abraham and Box (1978) versus the best model in Table 13.4.2, is calculated using [13.3.9] to be 0.95%.

13.5 FORECASTING AND SIMULATING WITH DESEASONALIZED MODELS

As noted in the introduction in Section 13.1, deseasonalized models can be easily used for forecasting and simulation by slightly extending the approaches described in Chapters 8 and 9, respectively. The key difference in the procedures is that one has to take into account the effects of deseasonalization when forecasting or simulating with the deseasonalized model.

Table 13.4.2. AIC values for ARMA models fitted to the deseasonalized ozone data.

ARMA Model	Box-Cox Transformation Parameter λ	F_μ	F_σ	AIC
(1,0)	1	2	1	-8.59
(1,0)	1	2	0	-8.19
(0,1)	1	2	1	-7.71
(0,1)	1	2	0	-7.43
(1,0)	0.5	2	1	-12.09
(1,0)	0.5	2	0	-10.27
(0,1)	0.5	2	1	-11.32
(0,1)	0.5	2	0	-9.55

Figure 13.5.1 outlines the procedure for forecasting with a deseasonalized model which is fitted to a seasonal time series. Firstly, the ARMA model fitted to the deseasonalized transformed series is used to obtain minimum mean square error forecasts of this data set. The procedure of Section 8.2.4 can be used for accomplishing this. Next, the forecasts for $z_{r,m}^{(\lambda)}$ are calculated using the inverse deseasonalization method which can be obtained from [13.2.2] or [13.2.3]. Thirdly, [13.2.1] can be used to determine the inverse Box-Cox transformation in order to obtain forecasts for $z_{r,m}$. The forecasts for $z_{r,m}$ are now in the same units as the original series to which the deseasonalized model was fitted. In Section 8.2.7, it is explained how one should correctly calculate minimum mean square error forecasts in the untransformed domain when the data were transformed using a Box-Cox transformation before fitting the model.

A similar procedure to that given in Figure 13.5.1 for forecasting can be used when simulating with a deseasonalized model. Simply replace the word forecast by simulate and forecasts by simulated values. In the first step, one of the procedures of Section 9.3 or 9.4 can be used to obtain the simulated sequences for the ARMA model fitted to the deseasonalized series. The rest of the procedure is the same as that explained for forecasting.

13.6 CONCLUSIONS

As exemplified by the two applications in Section 13.4, the procedures of Section 13.3 can be used to fit conveniently deseasonalized models to seasonal environmental time series. The two basic approaches to model construction presented in Sections 13.3.2 and 13.3.3 are summarized in Figure 13.3.1. For the case of the average monthly riverflows of the Saugeen River at Walkerton, Ontario, most of the Fourier components were needed to fully deseasonalize the series before fitting the ARMA model to the data. However, as shown by the monthly ozone data application in Section 13.4.3, a time series which does not possess a strongly pronounced seasonal structure may require less Fourier parameters for deseasonalization.

Deseasonalized models are designed for preserving the mean and variance within each season of the year. If one also wishes to capture the seasonal correlation structure, the periodic models described in the next chapter can be employed.

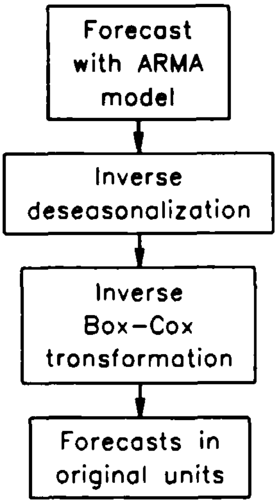

Forecast with ARMA model
Inverse deseasonalization
Inverse Box-Cox transformation
Forecasts in original units

Figure 13.5.1. Forecasting with a deseasonalized model.

PROBLEMS

13.1 In the SARIMA modelling approach to seasonal time series modelling in Chapter 12, differencing is employed to remove periodicity or seasonality. Within this chapter, deseasonalization is utilized for modelling periodicity. Employing some of the results of Kavvas and Delleur (1975) as well as other authors, compare the advantages and drawbacks of these two procedures for describing periodicity in time series.

13.2 Carry out an exploratory data analysis of an average monthly riverflow time series. Point out the main statistical characteristics of the data set. Of the three types of seasonal models presented in Part VI, which type of seasonal model do you think is most appropriate to fit to the data set?

13.3 Using the same monthly riverflow time series as in problem 13.2, fit the most appropriate fully deseasonalized model to the data set.

13.4 For the same time series used in problem 13.2, employ the approach of Section 13.3.3 to obtain a more parsimonious deseasonalized model. Comment upon your findings by comparing the results to those obtained in the two previous questions.

13.5 Examine appropriate graphs from an exploratory data analysis study of an average weekly time series of your choice. After explaining your exploratory findings, follow the procedure of Section 13.3.3 to construct a parsimonious model for fitting to the weekly series.

13.6 Execute problem 13.5 for an average daily time series that is of interest to you.

13.7 Fit the most appropriate deseasonalized model to a monthly riverflow time series. Follow the procedures of Sections 13.5 and 8.2.4 to forecast 24 steps ahead from the last data point in the time series. Also, plot the 90% probability interval.

13.8 Develop the most reasonable deseasonalized model to describe a monthly riverflow time series. Employing the techniques of Section 13.5 and Chapter 9, simulate and plot five sequences that are of the same length as the historical series. Compare the five simulated sequences and the historical data and then comment upon your findings.

REFERENCES

DATA SETS

Environment Canada (1977). Historical streamflow summary, Ontario. Water Survey of Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Canada.

Tiao, G. C., Box, G. E. P. and Hamming, W. J. (1973). Analysis of Los Angeles photochemical smog data: A statistical overview. Technical Report No. 331, Department of Statistics, University of Wisconsin, Madison, Wisconsin.

DESEASONALIZED MODELS

Abraham, B. and Box, G. E. P. (1978). Deterministic and forecast adaptive time dependent models. *Applied Statistics*, 27:120-130.

Croley II, T. E. and Rao, K. N. R. (1977). A manual for hydrologic time series deseasonalization and serial independence reduction. Technical report No. 199, lowa Institute of Hydraulic Research, The University of Iowa, Iowa City, Iowa.

Delleur, J. W., Tao, P. C. and Kavvas, M. L. (1976). An evaluation of the practicality and complexity of some rainfall and runoff time series models. *Water Resources Research*, 12(5):953-970.

Hipel, K. W. and McLeod, A. I. (1979). Modelling seasonal geophysical time series using deseasonalized models. Technical Report No. 52-XM-030579, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario.

Jayawardena, A. W. and Lai, F. (1989). Time series analysis of water quality data in Pearl River, China. *Journal of Environmental Engineering*, 115(3):590-607.

Kavvas, M. L. and Delleur, J. W. (1975). Removal of periodicities by differencing and monthly mean subtraction. *Journal of Hydrology*, 26:335-353.

McKerchar, A. I. and Delleur, J. W. (1974). Application of seasonal parametric linear stochastic models to monthly flow data. *Water Resources Research*, 10(2):246-255.

McMichael, F. C. and Hunter, J. S. (1972). Stochastic modeling of temperature and flow in rivers. *Water Resources Research*, 8(1):87-98.

Tao, P. C. and Delleur, J. W. (1976). Seasonal and nonseasonal ARMA models. *Journal of the Hydraulics Division, ASCE*, 102(HY10):1541-1559.

Thomas, H. A. and Fiering, M. B. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Maass, A., Hufshmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A. and Fair, M. G., Editors, *Design of Water Resources Systems*, 459-493. Harvard University Press, Cambridge, Massachusetts.

Yevjevich, V. and Harmancioglu, N. B. (1989). Description of periodic variation in parameters of hydrologic time series. *Water Resources Research*, 25(3):421-428.

TIME SERIES MODELLING

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716-723.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26:211-252.

CHAPTER 14

PERIODIC MODELS

14.1 INTRODUCTION

As emphasized by authors such as Moss and Bryson (1974), seasonal hydrological and other types of time series exhibit an autocorrelation structure which depends on not only the time lag between observations but also the season of the year. Furthermore, within a given season, usually second order stationarity is preserved by natural time series. For example, at a location in the northern hemisphere the monthly temperature for January across the years may fluctuate with constant variance around an overall mean of -5° C. In addition, the manner in which the January temperature is correlated with December and November as well as the previous January may tend to remain the same over the years. As another illustration of seasonally or periodically varying correlation, consider the case of runoff from snowmelt in late winter or early spring in a northern region. If the snowmelt is an important factor in runoff which might occur in either March or April, the correlation between observed riverflows for these months may be negative whereas at other times of the year it is usually positive. To model appropriately the foregoing and similar types of time series, periodic models can be employed. These models are ideal, for instance, for describing the average monthly flows of the Saugeen River at Walkerton, Ontario, Canada, plotted in Figure VI.1.

Two popular periodic models are the *PAR (periodic autoregressive)* and *PARMA (periodic ARMA) models*. When fitting a PAR model to a single seasonal series, a separate AR model is designed for each season of the year. In a similar manner, a PARMA model consists of having a separate ARMA model for each season of the year. Within hydrology, PAR modelling dates back to the research of Thomas and Fiering (1962) who proposed a specialized type of PAR model whereby the order of the AR operator for each season is fixed at unity.

Since the early 1960's a considerable amount of research has been executed in the area of periodic modelling. This research includes contributions by authors such as Gladyshev (1961, 1963), Jones and Brelsford (1967), Tao and Delleur (1976), Croley and Rao (1977), McLeod and Hipel (1978), Pagano (1978), Troutman (1979), Dunsmuir (1981), Tiao and Gruppe (1980), Sakai (1982), Salas et al. (1985), Cipra (1985a,b), Vecchia (1985a,b), Thompstone et al. (1985a), Cipra and Tlusty (1987), Jimenez et al. (1989) and McLeod (1993), as well as the books on stochastic hydrology referred to in Section 1.6.3.

As is explained in Section 14.3, a comprehensive range of *model construction* tools are available for conveniently fitting PAR models to seasonal time series. Because the theory and application of the PAR family of models are well-developed, this class of flexible models is stressed in this chapter. Nonetheless, some interesting developments in building PARMA models are pointed out in Section 14.7.

Subsequent to presenting model construction tools for use with PAR models in Section 14.3, a PAR model is developed for describing the average monthly flows of the Saugeen River plotted in Figure VI.1. A potential drawback of using a periodic model in an application is that the model often requires the use of a substantial number of parameters. Salas et al. (1980)

propose a Fourier series approach to reduce the number of model parameters in PAR and PARMA models. Thompstone et al. (1985a) suggest a procedure for combining individual AR models for various adjacent seasons, to obtain a single model for all of the seasons in the group. After joining appropriate seasons into groups, the overall periodic model that is fitted to the resulting data is called the *parsimonious periodic autoregressive (PPAR) model*. Subsequent to defining the PPAR model and presenting model construction methods in Section 14.5, PPAR models as well as other periodic models are fitted to seasonal hydrological time series in Section 14.6. Finally, in Section 14.8, *simulation experiments* are carried out to demonstrate that PAR and PPAR models statistically preserve *critical period statistics* which are used in reservoir design.

14.2 DEFINITIONS OF PERIODIC MODELS

14.2.1 Introduction

The definitions of PAR and PARMA models can be made from two different points of view. Firstly, PAR and PARMA models can be thought of as the periodic extensions of the nonseasonal AR and ARMA models, respectively, defined in Chapter 3. In other words, a PAR model consists of having a separate AR model for each season of the year whereas a PARMA model contains an ARMA model for each season. For both theoretical and practical reasons, the PAR and PARMA families of models are defined in these fashions in this chapter. For example, comprehensive model building procedures are now available for use with PAR models (Section 14.3) while significant progress has been made in developing model construction methods for employment with PARMA models (Section 14.7).

The second approach for defining PAR and PARMA models is to consider them to be special types of the multivariate ARMA models defined in Section 20.2. However, this approach is not recommended for various reasons. From an intuitive viewpoint, when one is trying to capture the physical characteristics of a natural phenomenon as portrayed in its time series of observations, it is more instructive and sensible to think of a periodic model as an extension of its nonseasonal counterpart. Hence, one can separately build models for each season of the year and then join them together to create the overall periodic model. Also, one can demonstrate theoretically that PAR and PARMA models can be written as equivalent multivariate AR and ARMA models, respectively, defined in Section 20.2. Conversely, multivariate AR and ARMA models can be represented as PAR and PARMA models, respectively.

The PAR family of models and some associated theoretical properties are presented in the next section. Following this, PARMA models are defined in Section 14.2.3.

14.2.2 PAR Models

Definition

For convenience, an observation in a time series is written in the same way as it is in Section 13.2.2 for deseasonalized models. When one is considering a time series having s seasons per year ($s = 12$ for monthly data) over a period of n years, let $z_{r,m}$ represent a time series observation in the rth year and mth season where $r = 1,2, \ldots, n$, and $m = 1,2, \ldots, s$. If required, the given data may be transformed by the Box-Cox transformation in [13.2.1] to form the transformed series denoted by $z_{r,m}^{(\lambda)}$. The purpose of the Box-Cox transformation is to correct problems such as heteroscedasticity and/or non-normality in the residuals of the PAR or PARMA model fitted to the time series.

In essence, a PAR model is formed by defining an AR model for each season of the year. The *PAR model* of order $(p_1, p_2, \ldots, p_s)$ is defined for season m as

$$z_{r,m}^{(\lambda)} - \mu_m = \sum_{i=1}^{p_m} \phi_i^{(m)} (z_{r,m-i}^{(\lambda)} - \mu_{m-i}) + a_{r,m} \qquad [14.2.1]$$

where μ_m is the mean of the series $z_{r,m}^{(\lambda)}$ for the mth season, $\phi_i^{(m)}$ is the AR coefficient for season m and ith lag, and $a_{r,m}$ is the innovation or white noise disturbance. The innovation series $a_{r,m}$ where $r = 1,2, \ldots, n$, is assumed to have an expected value of zero and a covariance defined by

$$cov(a_{r,m}, a_{r,m-i}) = \begin{cases} \sigma_m^2, & i = 0, \\ 0, & i \neq 0 \text{ for } i = 1,2, \ldots, s \end{cases} \qquad [14.2.2]$$

Hence, the $a_{r,m}$ disturbances are distributed as IID$(0, \sigma_m^2)$. By utilizing the backshift operator B, where $B^k z_{r,m}^{(\lambda)} = z_{r,m-k}^{(\lambda)}$, the model in [14.2.1] can be more succinctly written as

$$\phi^{(m)}(B)(z_{r,m}^{(\lambda)} - \mu_m) = a_{r,m}, \quad m = 1,2, \ldots, s \qquad [14.2.3]$$

where

$$\phi^{(m)}(B) = 1 - \phi_1^{(m)} B - \phi_2^{(m)} B^2 - \cdots - \phi_{p_m}^{(m)} B^{p_m}$$

is the AR operator of order p_m for season m in which $\phi_i^{(m)}$ is the ith AR parameter. For stationarity in season m, the roots of the seasonal characteristic equation $\phi^{(m)}(B) = 0$ must lie outside the unit circle. A necessary and sufficient condition for stationarity for a PAR model is given in [14.2.26].

Some authors recommend deseasonalizing the data using [13.2.3] before fitting a PAR or PARMA model to the time series [see, for example, Tao and Delleur (1976) and Croley and Rao (1977)]. However, when using the PAR model in [14.2.1] or [14.2.3], this step can easily be shown to be unnecessary, thereby reducing the number of model parameters. For example, suppose for the mth season that only one AR parameter were required and hence $p_m = 1$. From [14.2.1] or [14.2.3], this model is written as

$$z_{r,m}^{(\lambda)} - \mu_m = \phi_1^{(m)} (z_{r,m-1}^{(\lambda)} - \mu_{m-1}) + a_{r,m} \qquad [14.2.4]$$

which can be equivalently given as

$$\frac{z_{r,m}^{(\lambda)} - \mu_m}{\sqrt{\gamma_0^{(m)}}} = \phi_1^{(m')} \left[\frac{z_{r,m-1}^{(\lambda)} - \mu_{m-1}}{\sqrt{\gamma_0^{(m-1)}}} \right] + a'_{r,m} \qquad [14.2.5]$$

where

$$\gamma_0^{(m)} = var(z_{r,m}^{(\lambda)}), \quad \text{for } m = 1,2, \ldots, s;$$

$$\phi_1^{(m')} = \left[\frac{\gamma_0^{(m-1)}}{\gamma_0^{(m)}} \right]^{0.5} \phi_1^{(m)};$$

and

$$a'_{r,m} = \left[\frac{\gamma_0^{(m-1)}}{\gamma_0^{(m)}} \right]^{0.5} a_{r,m} .$$

Stationarity

In Chapter 20, the general multivariate ARMA model is defined. As explained by authors such as Rose (1977), Newton (1982), Vecchia (1985a,b), Obeysekera and Salas (1986), Haltiner and Salas (1988), Bartolini et al. (1988) and Ula (1990), a PAR model can be equivalently written as a special case of the multivariate ARMA model. Because the stationarity conditions for a multivariate ARMA are known, they are also available for a PAR model. Consider, for example, the case of a PAR model in [14.2.4] for which there is one AR parameter for each of the s seasons. The stationarity requirement for this model is (Obeysekera and Salas, 1986)

$$\left| \prod_{m=1}^{s} \phi_1^{(m)} \right| < 1 \qquad [14.2.6]$$

Periodic Autocorrelation Function

The theoretical ACF for the PAR model in [14.2.1] or [14.2.3] for season m can be found by following a similar procedure to that used for obtaining the theoretical ACF for the nonseasonal AR model in Section 3.2.2. First, however, it is necessary to formulate some definitions. For season m, the theoretical *periodic autocovariance function* at lag k is defined for $z_{r,m}^{(\lambda)}$ as

$$\gamma_k^{(m)} = E[(z_{r,m}^{(\lambda)} - \mu_m)(z_{r,m-k}^{(\lambda)} - \mu_{m-k})] \qquad [14.2.7]$$

for $m = 1,2, \ldots, s$, where μ_m and μ_{m-k} are the theoretical means for seasons m and $m-k$, respectively. When $k = 0$, the periodic autocovariance is simply the variance, $\gamma_0^{(m)}$, of the random variable representing the observations in season m.

A standardized variable that is more convenient to deal with than $\gamma_k^{(m)}$, is the theoretical *periodic ACF* which is defined for season m at lag k as

$$\rho_k^{(m)} = \frac{\gamma_k^{(m)}}{\sqrt{\gamma_0^{(m)}\gamma_0^{(m-k)}}} \qquad [14.2.8]$$

Due to the form of [14.2.8], the theoretical periodic ACF is dimensionless and, hence, independent of the scale of measurement. Furthermore, the possible values of $\rho_k^{(m)}$ range from -1 to 1, where $\rho_k^{(m)}$ has a magnitude of unity at lag zero.

Given the above definitions of periodic linear dependence, one can find the theoretical periodic ACF for the PAR model in [14.2.1] or [14.2.3]. For season m, multiply [14.2.1] by $z_{r,m-k}^{(\lambda)} - \mu_{m-k}$ and take expected values to obtain

$$\gamma_k^{(m)} = \phi_1^{(m)}\gamma_{k-1}^{(m-1)} + \phi_2^{(m)}\gamma_{k-2}^{(m-2)} + \cdots + \phi_{p_m}^{(m)}\gamma_{k-p_m}^{(m-p_m)} + E[(z_{r,m-k}^{(\lambda)} - \mu_{m-k})a_{r,m}] \quad [14.2.9]$$

for $k \geq 0$ and $m = 1,2,\ldots,s$. The last term on the right hand side of [14.2.9] is zero for $k > 0$ because $z_{r,m-k}^{(\lambda)}$ is only a function of the disturbances $a_{r,m}$ up to time $m-k$ and $a_{r,m}$ is independent of these shocks. Hence, for $k > 0$ [14.2.9] becomes

$$\gamma_k^{(m)} = \phi_1^{(m)}\gamma_{k-1}^{(m-1)} + \phi_2^{(m)}\gamma_{k-2}^{(m-2)} + \cdots + \phi_{p_m}^{(m)}\gamma_{k-p_m}^{(m-p_m)} \quad [14.2.10]$$

By using the periodic AR operator given in [14.2.3], one can rewrite [14.2.10] for season m as

$$\phi^{(m)}(B)\gamma_k^{(m)} = 0 \quad \text{for } k > 0 \quad [14.2.11]$$

where B operates on the subscript k and the superscript (m) in $\gamma_k^{(m)}$. The relationship in [14.2.11] is valid for each season $m = 1,2,\ldots,s$. Because of the form of [14.2.10] and [14.2.11], the theoretical autocovariance function attenuates for a PAR process in season m when $p_m > 0$.

Periodic Yule-Walker Equations

Following the approach used for a nonseasonal AR model in Section 3.2.2, one can find the theoretical Yule-Walker equations for a PAR model. Specifically, by setting $k = 1,2,\ldots,p_m$, in [14.2.10], one obtains the *periodic Yule-Walker equations* for season m as:

$$\begin{array}{ccccccccc}
\gamma_1^{(m)} & = & \phi_1^{(m)}\gamma_0^{(m-1)} & + & \phi_2^{(m)}\gamma_1^{(m-2)} & + & \cdots & + & \phi_{p_m}^{(m)}\gamma_{p_m-1}^{(m-p_m)} \\
\gamma_2^{(m)} & = & \phi_1^{(m)}\gamma_1^{(m-1)} & + & \phi_2^{(m)}\gamma_0^{(m-2)} & + & \cdots & + & \phi_{p_m}^{(m)}\gamma_{p_m-2}^{(m-p_m)} \\
\cdot & \cdot & \cdot & \cdot & \cdot & & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & & \cdots & & \cdot \\
\gamma_{p_m}^{(m)} & = & \phi_1\gamma_{p_m-1}^{(m-1)} & + & \phi_2^{(m)}\gamma_{p_m-2}^{(m-2)} & + & \cdots & + & \phi_{p_m}^{(m)}\gamma_0^{(m-p_m)}
\end{array} \quad [14.2.12]$$

By writing the periodic Yule-Walker equations in [14.2.12] in matrix form, the relationship for expressing the AR parameters for season m is

$$\underline{\phi}^{(m)} = \left[\underline{\Gamma}_{p_m}^{(m)}\right]^{-1}\underline{\gamma}^{(m)} \quad [14.2.13]$$

where

$$\underline{\phi}^{(m)} = \begin{bmatrix} \phi_1^{(m)} \\ \phi_2^{(m)} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{p_m}^{(m)} \end{bmatrix}, \underline{\gamma}^{(m)} = \begin{bmatrix} \gamma_1^{(m)} \\ \gamma_2^{(m)} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_m^{(m)} \end{bmatrix}, \underline{\Gamma}_{p_m}^{(m)} = \begin{bmatrix} \gamma_0^{(m-1)} & \gamma_1^{(m-2)} & \cdots & \gamma_{p_m-1}^{(m-p_m)} \\ \gamma_1^{(m-1)} & \gamma_0^{(m-2)} & \cdots & \gamma_{p_m-2}^{(m-p_m)} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot \quad \cdots & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \gamma_{p_m-1}^{(m-1)} & \gamma_{p_m-2}^{(m-2)} & \cdots & \gamma_0^{(m-p_m)} \end{bmatrix}$$

By setting $k = 0$ in [14.2.9], the expression for the variance $\gamma_0^{(m)}$ is

$$\gamma_0^{(m)} = \phi_1^{(m)}\gamma_1^{(m-1)} + \phi_2^{(m)}\gamma_2^{(m-2)} + \cdots + \phi_{p_m}^{(m)}\gamma_{p_m}^{(m-p_m)} + \sigma_m^2 \qquad [14.2.14]$$

where $E[(z_{r,m}^{(\lambda)} - \mu_m)(a_{r,m})] = \sigma_m^2$ since $z_{r,m}^{(\lambda)}$ is only correlated with $a_{r,m}$ due to the most recent shock $a_{r,m}$. As is explained in Section 14.3.3, the periodic Yule-Walker equations in [14.2.12] or [14.2.13] provide a means of obtaining efficient moment estimates for the parameters of the PAR model in [14.2.1] or [14.2.3].

Periodic Partial Autocorrelation Function

Since the periodic autocorrelation function of a PAR model in season m for which $p_m > 0$ attenuates and does not truncate at a specified lag, it would be useful for identification purposes to define a function which cuts off. To accomplish this one can define the periodic PACF for a PAR model in a manner similar to that done in Section 3.2.2 for a nonseasonal AR model.

For season m, the *periodic PACF* is defined as the last AR parameter of an AR model of order p_m. Therefore, in the Yule-Walker equations in [14.2.12], $\phi_{p_m}^{(m)}$ is by definition the periodic PACF at lag p_m. By setting p_m to values of 1,2, . . . , in [14.2.12], one can define the periodic PACF in season m for lags 1,2, . . . , respectively. Because of the definition of the theoretical periodic PACF, it must be equal to zero after lag p_m in season m when the order of the AR model in this season is p_m. Furthermore, the possible values of the theoretical PACF fall between -1 and +1.

Markov Model

For a Markov model in season m the order is $p_m = 1$. A Markov model for season m is written in [14.2.4]. When the PAR is Markov for each of the s seasons, the stationarity condition for the overall Markov PAR model is the one given in [14.2.6].

The periodic Yule-Walker equations for a PAR model are written in [14.2.12]. By setting $\phi_2^{(m)}$ to $\phi_p^{(m)}$ equal to zero, this equation becomes

$$\gamma_1^{(m)} = \phi_1^{(m)}\gamma_0^{(m-1)}$$

$$\gamma_2^{(m)} = \phi_1^{(m)}\gamma_1^{(m-1)}$$

$$\gamma_3^{(m)} = \phi_1^{(m)}\gamma_2^{(m-1)}$$

In general,

$$\gamma_k^{(m)} = \phi_1^{(m)}\gamma_{k-1}^{(m-1)}$$

Hence, the theoretical periodic autocovariance function attenuates for increasing lag k. However, by definition the theoretical periodic PACF cuts off and is exactly equal to zero after lag one for a Markov model.

14.2.3 PARMA Models

Definition

As is also the case for the PAR model in Section 14.2.2, let $z_{r,m}^{(\lambda)}$ be an observation in the rth year and mth season for $r = 1,2,\ldots,n$, and $m = 1,2,\ldots,s$, where the exponent λ indicates that the observation may be transformed using the Box-Cox transformation in [13.2.1]. A PARMA model is created by defining a separate ARMA model for each season of the year. The PARMA model of order $(p_1,q_1;p_2,q_2;\cdots;p_s,q_s)$ is defined for season m as

$$\phi^{(m)}(B)(z_{r,m}^{(\lambda)} - \mu_m) = \theta^{(m)}(B)a_{r,m}, \quad m = 1,2,\ldots,s \qquad [14.2.15]$$

where μ_m is the mean for series $z_{r,m}^{(\lambda)}$ for the mth season, $\phi^{(m)}(B) = 1 - \phi_1^{(m)}B - \phi_2^{(m)}B^2 - \cdots - \phi_{p_m}^{(m)}B^{p_m}$, is the AR operator of order p_m for season m in which $\phi_i^{(m)}$ is the ith AR parameter, and $\theta^{(m)}(B) = 1 - \theta_1^{(m)}B - \theta_2^{(m)}B^2 - \cdots - \theta_{q_m}^{(m)}B^{q_m}$, is the MA operator of order q_m for season m in which $\theta_i^{(m)}$ is the ith MA parameter. The innovation series $a_{r,m}$ where $r = 1,2,\ldots,n$, for each m is assumed to be distributed as IID$(0,\sigma_m^2)$ which is the same as that for the PAR model in [14.2.1].

Using the AR and MA operators to define the PARMA model in [14.2.15] provides an economical and convenient format for writing this model. Also, the operator format in [14.2.15] can be easily manipulated for mathematical purposes. Nonetheless, one could also write the PARMA model for season m without the operator notation as

$$z_{r,m}^{(\lambda)} - \mu_m = \sum_{i=1}^{p_m}\phi_i^{(m)}(z_{r,m-i} - \mu_{m-i}) + a_{r,m} - \sum_{i=1}^{q_m}\theta_i^{(m)}a_{r,m-i}, \quad m = 1,2,\ldots,s \qquad [14.2.16]$$

Stationarity and Invertibility

The PARMA model given in [14.2.15] can be equivalently written as a particular case of the general multivariate ARMA model presented in Chapter 20. Since the stationarity and invertibility conditions for the general multivariate ARMA model are available, they are, of course, also known for the PARMA model (Rose, 1977; Vecchia, 1985a,b; Obeysekera and Salas, 1986; Bartolini et al., 1988; Ula, 1990). As an example of how these conditions are written for a specific PARMA model, consider a PARMA model from [14.2.15] for which there is one AR and one MA parameter for each of the m seasons. The stationarity restriction for this model is given in [14.2.6] while the invertibility requirement is

$$\left|\prod_{m=1}^{s}\theta_1^{(m)}\right| < 1 \qquad [14.2.17]$$

Periodic Autocorrelation Function

In Section 3.4.2, it is explained how the theoretical autocovariance function or, equivalently, the theoretical ACF can be determined for a nonseasonal ARMA(p,q) model. A similar approach can be followed to derive the system of equations for solving for the periodic autocovariance function in [14.2.7] for a PARMA model.

The steps required for accomplishing this are now described. For a given season m, multiply both sides of [14.2.15] by $z_{r,m-k}^{(\lambda)} - \mu_{m-k}$ and take the expected values to obtain

$$\gamma_k^{(m)} - \phi_1^{(m)}\gamma_{k-1}^{(m-1)} - \phi_2^{(m)}\gamma_{k-2}^{(m-2)} - \cdots - \phi_{p_m}^{(m)}\gamma_{k-p_m}^{(m-p_m)}$$
$$= \gamma_{za}^{(m)}(k) - \theta_1^{(m)}\gamma_{za}^{(m-1)}(k-1) - \cdots - \theta_{q_m}^{(m)}\gamma_{za}^{(m-q_m)}(k-q_m) \qquad [14.2.18]$$

where $\gamma_k^{(m)}$ is the theoretical periodic autocovariance function in [14.2.7] and

$$\gamma_{za}^{(m)}(k) = E[(z_{r,m-k}^{(\lambda)} - \mu_{m-k})a_{r,m}] \qquad [14.2.19]$$

is the cross covariance function between $z_{r,m-k}^{(\lambda)} - \mu_{m-k}$ and $a_{r,m}$. Since $z_{r,m-k}^{(\lambda)}$ is only dependent upon shocks which have occurred up to time $(r,m-k)$, it follows that

$$\begin{aligned} &\gamma_{za}^{(m)}(k) = 0, \quad k > 0 \\ &\gamma_{za}^{(m)}(k) \neq 0, \quad k \leq 0 \end{aligned} \qquad [14.2.20]$$

Because of the $\gamma_{za}^{(m)}(k)$ terms in [14.2.18], one must derive other relationships before one can solve for the periodic autocovariances. This can be carried out by multiplying [14.2.15] by $a_{r,m-k}$ and taking expectations to obtain

$$\gamma_{za}^{(m-k)}(-k) - \phi_1^{(m)}\gamma_{za}^{(m-k)}(-k+1) - \phi_2^{(m)}\gamma_{za}^{(m-k)}(-k+2) - \cdots - \phi_{p_m}^{(m)}\gamma_{za}^{(m-k)}(-k+p_m)$$
$$= -[\theta_k^{(m)}]\sigma_m^2 \qquad [14.2.21]$$

where

$$[\theta_k^{(m)}] = \begin{cases} \theta_k^{(m)}, & k = 1,2,\ldots,q_m \\ -1, & k = 0 \\ 0, & otherwise \end{cases}$$

and $E[a_{r,m}a_{r,m-k}]$ is as defined in [14.2.2].

Equations [14.2.18] and [14.2.21] can be employed to solve for the theoretical *periodic autocovariance function* for a PARMA model for each season. For $k > q_m$, equation [14.2.18] reduces to

$$\gamma_k^{(m)} - \phi_1^{(m)}\gamma_{k-1}^{(m-1)} - \phi_2^{(m)}\gamma_{k-2}^{(m-2)} - \cdots - \phi_{p_m}^{(m)}\gamma_{k-p_m}^{(m-p_m)} = 0$$

or

$$\phi_m^{(m)}(B)\gamma_k^{(m)} = 0 \qquad [14.2.22]$$

where the differencing operator B operates on both the subscript and superscript in $\gamma_k^{(m)}$. If $k > \max(p_m,q_m)$, then [14.2.22] can be used to calculate the $\gamma_k^{(m)}$ directly from previous values. For $k = 0,1,2,\ldots,\max(p_m,q_m)$, equation [14.2.21] can be employed for solving for the periodic cross covariance function $\gamma_{za}^{(m)}(k)$ which can be substituted into [14.2.18] in order to solve for the periodic autocovariance function for the $z_{r,m}^{(\lambda)}$. By employing [14.2.8], one can easily calculate the theoretical periodic ACF after determining the theoretical periodic autocovariance function.

Recall that for a nonseasonal ARMA model in Section 3.4.2, the theoretical autocovariance function or theoretical ACF attenuates for increasing values of lag k. In a similar fashion, one can see from the form of [14.2.20] that the theoretical periodic autocovariance function dies off for a PARMA model in which $p_m \neq 0$ in season m.

Periodic Partial Autocorrelation Function

For season m, the PARMA model in [14.2.15] can be written as an infinite AR model by writing it as

$$a_{r,m} = \theta^{(m)}(B)^{-1}\phi^{(m)}(B)(z_{r,m}^{(\lambda)} - \mu_m) \qquad [14.2.23]$$

where $\theta^{(m)}(B)^{-1}$ is an infinite series in B for $q_m \geq 1$. Because the definition of the theoretical *periodic PACF* is based upon an AR process, the periodic PACF is infinite in extent for a PARMA model and dies off with increasing lag. At higher lags, the behaviour of the periodic PACF depends upon the MA parameters and is dominated by a combination of damped exponentials and/or damped sine waves.

Three Formulations of a PARMA Model

In Section 3.4.3, it is explained how a nonseasonal ARMA model can be expressed in three equivalent forms. These same three formats can also be used with a PARMA model in season m. One formulation is to use the difference equation given in [14.2.15]. A second technique is to write the model as a pure MA model, which is also called the *random shock form*. Finally, by formulating the model as a pure AR model one obtains the inverted form for the model.

In random shock form, the PARMA model for season m is written as

$$\begin{aligned} z_{r,m}^{(\lambda)} - \mu_m &= \phi^{(m)}(B)^{-1}\theta^{(m)}(B)a_{r,m} \\ &= a_{r,m} + \psi_1^{(m)}a_{r,m-1} + \psi_2^{(m)}a_{r,m-2} + \cdots \\ &= a_{r,m} + \psi^{(m)}Ba_{r,m} + \psi_2^{(m)}B^2a_{r,m} + \cdots \\ &= (1 + \psi_1^{(m)}B + \psi_2^{(m)}B^2 + \cdots)a_{r,m} \\ &= \psi^{(m)}(B)a_{r,m} \end{aligned} \qquad [14.2.24]$$

where $\psi^{(m)}(B) = 1 + \psi_1^{(m)}B + \psi_2^{(m)}B^2 + \cdots$, is the random shock or infinite MA operator for season m and $\psi_i^{(m)}$ is the ith parameter, coefficient or weight of $\psi^{(m)}(B)$. There are a variety of reasons for expressing a model in random shock form. For example, when forecasting in season m the $\psi_i^{(m)}$ weights are needed to calculate the variance of the forecasts (see [8.2.13] for the case of an ARMA model). When simulating in season m using a PARMA model, one way to simulate data is to write the model in random shock form and then to use this structure for producing the synthetic sequences (see Section 9.3 for the case of an ARMA model). Finally, by writing PARMA models in random shock form, the magnitude and sign of the $\psi_i^{(m)}$ parameters can be compared across models.

Following the arguments given in Section 3.4.3 to develop [3.4.21] for an ARMA model, one can obtain the $\psi_k^{(m)}$ weights from the $\phi_k^{(m)}$ and $\theta_k^{(m)}$ parameters for a PARMA model in season m by utilizing the expression

$$\phi^{(m)}(B)\psi_k^{(m)} = -\theta_k^{(m)} \qquad [14.2.25]$$

where B operates on k, $\psi_0^{(m)} = 1$, $\psi_k^{(m)} = 0$ for $k < 0$ and $\theta_k^{(m)} = 0$ if $k > q$. Rules for deciding upon how many random shock parameters to calculate and examples for determining these parameters are given in Section 3.4.3.

Because a PAR model is a special case of a PARMA model, one can, of course, write a PAR for season m in the random shock form given in [14.2.4]. As shown by Troutman (1979), a necessary and sufficient condition for periodic stationarity for a PAR model is

$$\sum_{i=0}^{\infty}(\psi_i^{(m)})^2 < \infty, \quad m = 1,2,\ldots,s \qquad [14.2.26]$$

To express the PARMA model in season m in *inverted form*, equation [14.2.15] is rewritten as

$$\begin{aligned}
a_{r,m} &= \theta^{(m)}(B)^{-1}\phi^{(m)}(B)(z_{r,m}^{(\lambda)} - \mu_m) \\
&= (z_{r,m}^{(\lambda)} - \mu_m) - \Pi_1^{(m)}(z_{r,m-1}^{(\lambda)} - \mu_{m-1}) - \Pi_2^{(m)}(z_{r,m-2}^{(\lambda)} - \mu_{m-2}) - \cdots \\
&= (z_{r,m} - \mu_m) - \Pi_1^{(m)}B(z_{r,m}^{(\lambda)} - \mu_m) - \Pi_2^{(m)}B^2(z_{r,m}^{(\lambda)} - \mu_m) - \cdots \\
&= (1 - \Pi_1^{(m)}B - \Pi_2^{(m)}B^2 - \cdots)(z_{r,m}^{(\lambda)} - \mu_m) \\
&= \Pi^{(m)}(B)(z_{r,m}^{(\lambda)} - \mu_m) \qquad [14.2.27]
\end{aligned}$$

where $\Pi^{(m)}(B) = 1 - \Pi_1^{(m)}B - \Pi_2^{(m)}B^2 - \cdots$, is the inverted or infinite AR operator for season m and $\Pi_i^{(m)}$ is the ith parameter, coefficient or weight of $\Pi^{(m)}(B)$. By comparing [14.2.26] and [14.2.24], one can see that

$$\psi^{(m)}(B)^{-1} = \Pi^{(m)}(B) \qquad [14.2.28]$$

Given the seasonal AR and MA parameters, one may wish to determine the inverted parameters. To achieve this, one can use the expression

$$\theta^{(m)}(B)\Pi_k^{(m)} = \phi_k^{(m)} \qquad [14.2.29]$$

where B operates on k, $\Pi_0^{(m)} = -1$, $\Pi_k^{(m)} = 0$ for $k < 0$, and $\phi_k^{(m)} = 0$ if $k > p$. Except for notational differences, [14.2.28] is the same as [3.4.27] which is used for obtaining the inverted weights for a nonseasonal ARMA model. Representative examples for calculating the inverted weights are presented in Section 3.4.3.

Example of a PARMA Model

If $p_m = q_m = 1$ for season m, a PARMA model for that season is written following [14.2.15] as

$$(1-\phi_1^{(m)}B)(z_{r,m}-\mu_m)=(1-\theta_1^{(m)}B)a_{r,m} \qquad [14.2.30]$$

To obtain the theoretical autocovariance function, one must solve [14.2.18] and [14.2.21] after setting all AR and MA parameters equal to zero except for $\phi_1^{(m)}$ and $\theta_1^{(m)}$. The reader can refer to Section 3.4.3 for examples of how to calculate the random shock and inverted parameters for a nonseasonal ARMA(1,1) model. The same approaches can be used for a PARMA model with $p_m = q_m = 1$ by replacing ϕ_i, θ_i, ψ_i, and Π_i by $\phi_i^{(m)}$, $\theta_i^{(m)}$, $\psi_i^{(m)}$ and $\Pi_i^{(m)}$ in Section 3.4.3.

14.3 CONSTRUCTING PAR MODELS

14.3.1 Introduction

As noted in Section 14.1, model construction techniques for PAR models are highly developed. Indeed, as demonstrated by research referenced in Section 14.1, PAR models can be conveniently used in practical applications and produce useful results. Consequently, this section concentrates upon how to construct PAR models by following the three stages of model construction. Applications for clearly illustrating how the construction techniques for PAR modelling are implemented in practice are presented in Sections 14.5 and 14.6 as well as Chapter 15. Finally, model construction methods for PPAR and PARMA models are given in Sections 14.5.3 and 14.7, respectively.

14.3.2 Identifying PAR Models

Introduction

Thomas and Fiering (1962) originally suggested that one could fit PAR models of order one for each season to monthly hydrological time series. More recently, authors such as Salas et al. (1980) and Thompstone et al. (1985a,b) have suggested that the order of the AR operator for each season be properly identified. Based upon the results of an extensive forecasting study, Noakes et al. (1985) recommend that the best way to identify a PAR model is to employ the periodic ACF and PACF. Consequently, this approach to designing a PAR model is explained in this section. Another identification method which uses the AIC in conjunction with subset autoregression and the algorithm of Morgan and Tatar (1972) is outlined in Section 14.3.3. Moreover, two procedures for efficiently estimating the parameters of PAR models are described in Section 14.3.3 while diagnostic checks are discussed in Section 14.3.4. Finally, the results of the forecasting study of Noakes et al. (1985) are presented in Section 15.3 to demonstrate that PAR models identified using the periodic ACF and PACF forecast better than PAR models designed using other approaches as well as the deseasonalized and SARIMA models of Chapters 13 and 12, respectively.

Sample Periodic ACF: The theoretical periodic autocovariance function and ACF at lag k for the series $z_{r,m}^{(\lambda)}$ are defined in [14.2.7] and [14.2.8], respectively. In a practical application, the theoretical variables used in these equations are estimated using the sample time series $z_{r,m}$, where the years $r = 1,2,\ldots,n$, and the seasons $m = 1,2,\ldots,s$. To rectify problems with non-normality and/or heteroscedasticity in the residuals of the fitted PAR model, often the original series, $z_{r,m}$, is transformed using the Box-Cox transformation in [13.2.1] to obtain the transformed series $z_{r,m}^{(\lambda)}$. The theoretical variables in [14.2.7] and [14.2.8] are then estimated for the $z_{r,m}^{(\lambda)}$ series. More specifically, for the mth season, the mean, μ_m, is estimated using

$$\hat{\mu}_m = \frac{1}{n}\sum_{r=1}^{n} z_{r,m}^{(\lambda)} \qquad [14.3.1]$$

where $m = 1,2,\ldots,s$. To estimate the theoretical periodic autocovariance function, $\gamma_k^{(m)}$, in [14.2.7] for lag k and season m, the following formula is utilized:

$$c_k^{(m)} = \frac{1}{n}\sum_{r=1}^{n}(z_{r,m}^{(\lambda)} - \hat{\mu}_m)(z_{r,m-k}^{(\lambda)} - \mu_{m-k}) \qquad [14.3.2]$$

for $m = 1,2,\ldots,s$. When the lag k is zero, one obtains the estimate of the variance of the observations in season m, which is given as:

$$c_o^{(m)} = \frac{1}{n}\sum_{r=1}^{n}(z_{r,m}^{(\lambda)} - \hat{\mu}_m)^2, \quad m = 1,2,\ldots,s. \qquad [14.3.3]$$

The *sample or estimated theoretical periodic ACF* at lag k is determined for $\rho_k^{(m)}$ using

$$r_k^{(m)} = \frac{c_k^{(m)}}{\sqrt{c_o^m c_o^{m-k}}} \qquad [14.3.4]$$

where $m = 1,2,\ldots,s$.

Because the periodic ACF is symmetric about lag zero, it is only necessary to plot the sample ACF for season m from lag one to a maximum lag of about $n/4$. A separate sample ACF graph is made for each season of the year. To ascertain which values of the estimated ACF for period or season m are significantly different from zero, the approximate 95% confidence interval can be plotted. The sample ACF is asymptotically distributed as NID$(0,\frac{1}{n})$ at any lag. Consequently, the approximate 95% confidence interval is $\pm 1.96\sqrt{n}$.

As explained in Section 14.2.2, the theoretical ACF for a PAR model in season m, attenuates if AR parameters are in the model. Consequently, if the sample periodic ACF dies off for season m, this indicates that one or more AR parameters are needed in this season for the PAR model which is fitted to the series. If no values of the sample periodic ACF are significantly different from zero, this means that one can model this season using white noise by setting $p_m = 0$ in the PAR model in [14.2.3].

Sample Periodic PACF: For a given seasonal time series, the periodic PACF can be determined for each season of the year. The definition for the periodic PACF is derived from the definition of the PAR model. In particular, assuming that the AR model for season m is of order p_m, the PACF for that season is $\phi_{p_m}^{(m)}$. Be setting $p_m = 1,2,\ldots,$ the PACF is defined for lags 1,2,... .

For the case of a nonseasonal time series, one uses the Yule-Walker equations in [3.2.12] or [3.2.17] to estimate the PACF. Likewise, for the situation of a periodic or seasonal time series one can utilize the periodic Yule-Walker equations in [14.2.12] or [14.2.13] to estimate the periodic PACF.

To obtain Yule-Walker estimates for the AR parameters for season m in the PAR model in [14.2.1] or [14.2.3], simply replace each $\gamma_k^{(m)}$ in [14.2.12] or [14.2.13] by its estimate $c_k^{(m)}$ from [14.3.2]. This can be carried out for each of the s seasons by estimating the $\phi_i^{(m)}$ for $i = 1,2,\ldots,p_m$, for each season $m = 1,2,\ldots,s$. The resulting estimated PAR model is periodic stationary (Troutman, 1979) and the estimates are asymptotically efficient (Pagano, 1978). Furthermore, the estimates corresponding to different seasons are asymptotically independent. Sakai (1982) presents a practical computational algorithm for estimating the periodic AR parameters and, hence, also the periodic PACF from the periodic Yule-Walker equations in [14.2.12] or [14.2.13].

For period or season m, the correct order for a PAR model is given as p_m in [14.2.1]. Sakai (1982) shows that the sample PACF for a given season is asymptotically distributed as $\text{NID}(0,\frac{1}{n})$ at any lag greater than p_m. Therefore, the 95% confidence interval is $\pm 1.96\sqrt{n}$. The sample PACF and approximate 95% confidence interval can be plotted for each season up to a maximum lag of about $\frac{n}{4}$.

By definition the theoretical PACF in season m cuts off after lag p_m to a value of exactly zero. Consequently, if the sample periodic PACF is not significantly different from zero after lag p_m, this indicates that the order of the AR model fitted to the series in season m should be p_m. If none of the values of the sample periodic PACF in season are significantly different from zero, the model for season m within the overall PAR model should be white noise. In this case, p_m is set equal to zero.

Periodic IACF and IPACF

In Section 5.3, the sample IACF and IPACF are recommended as additional identification tools for determining the orders of the AR and MA operators in a nonseasonal ARMA model. One can define the periodic versions of these functions for use in identifying the order of the AR model for each season of a PAR model.

For season m, the *theoretical periodic IACF* of a PARMA model is defined to be the ACF of a PARMA model having the AR and MA components of orders q_m and p_m, respectively (i.e. the AR and MA operators are exchanged with one another). The PACF of this process for season m is defined to be the *theoretical periodic IPACF*.

For a PAR model having an AR operator of order p_m in season m, the IACF truncates after lag p_m. Thus, the behaviour of the periodic IACF is similar to that of the periodic PACF. Likewise, the periodic IPACF mimics the behaviour of the periodic ACF. For both of these latter functions, their values die off for increasing lags in season m when $p_m \neq 0$.

Further research is required for obtaining efficient estimates for the sample periodic IACF and IPACF. One could, for example, adopt estimation procedures similar to those developed for the nonseasonal versions of these functions in Chapter 5.

Tests for Periodic Correlation

The sample periodic ACF and PACF provide a means for detecting periodic correlation in seasonal time series and also information for designing a PAR model to fit to the series. Other approaches for finding periodic correlation in a data set include the statistical tests described by Hurd and Gerr (1991) and Vecchia and Ballerini (1991). When periodic correlation is present, one, should, of course, fit a periodic model such as a PAR or PARMA model to the time series under consideration. Tiao and Gruppe (1980) discuss the negative consequences of not using an appropriate periodic model when the data possesses periodic correlation.

14.3.3 Calibrating PAR Models

Introduction

A major advantage of using PAR models in practical applications is that two good algorithms are available for estimating the parameters of PAR models. In particular, the two estimation methods described in the next two subsections are the Yule-Walker estimator and multiple linear regression. These two estimator techniques are efficient both from statistical and computational viewpoints.

For deciding upon the order of the AR operator in each season, one can use plots of the sample periodic ACF and PACF, as explained in Section 14.3.2. Additionally, one can employ the AIC which is derived for the case of PAR models in this section. Finally, it is explained how the algorithm of Morgan and Tatar (1972) can be used in conjunction with the AIC to select the order of each AR operator in a PAR model.

Periodic Yule-Walker Estimator

The technique for obtaining Yule-Walker estimates for the parameters of a PAR model is explained in Section 14.3.2 under the subsection entitled Sample Periodic PACF. Even though this method is in fact a moment estimator, it is still efficient statistically for use with PAR model.

As discussed in Section 14.3.2, the periodic Yule-Walker equation in [14.2.12] or [14.2.13] can be used to obtain the estimates of the parameters for each season. Each theoretical ACF is replaced by its sample estimate from [14.3.2] and then the algorithm of Sakai (1982) is used to estimate the AR parameters for each season m using the periodic Yule-Walker equations. The parameters for each season can be estimated separately and the parameter estimates are asymptotically efficient (Pagano, 1978). Furthermore, the calibrated PAR model is periodic stationary (Troutman, 1979).

Multiple Linear Regression

Although λ could be estimated, assume that it is fixed at some value such as $\lambda = 0.5$ or $\lambda = 0$ for a square root or natural logarithmic transformation, respectively. For season m, the mean parameter μ_m is estimated by

$$\hat{\mu}_m = \frac{1}{n}\sum_{r=1}^{n} z_{r,m}^{(\lambda)}, \quad m = 1,2,\ldots,s \qquad [14.3.5]$$

For season or period (m), let $\beta_m = \phi_1^{(m)},\phi_2^{(m)},\ldots,\phi_{p_m}^{(m)}$, denote the vector of AR parameters for the PAR model and $\beta_m = \hat{\phi}_1^{(m)},\hat{\phi}_2^{(m)},\ldots,\hat{\phi}_{p_m}^{(m)}$, stand for the vector of estimated parameters. An efficient conditional maximum likelihood estimate $\hat{\beta}_m$ of $\hat{\beta}_m$ is are obtained directly from the multiple linear regression of $z_{r,m}^{(\lambda)}$ on $z_{r,m-1}^{(\lambda)},z_{r,m-2}^{(\lambda)},\ldots,z_{r,m-p_m}^{(\lambda)}$.

The estimated innovations or residuals denoted as $\hat{a}_{r,m}$ are calculated from [14.2.3] by setting initial values to zero and the residual variance, σ_m^2 is then estimated by

$$\hat{\sigma}_m^2 = \frac{1}{n}\sum_{r=1}^{n}\hat{a}_{r,m}^2, \quad m = 1,2,\ldots,s \qquad [14.3.6]$$

Other Estimation Results

Pagano (1978) shows that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean zero and covariance matrix $\frac{1}{n}\mathbf{I}_m^{-1}$, where

$$\mathbf{I}_m = \frac{1}{\sigma_m^2}(\gamma_{i-j}^{(m)})$$

In practice, an estimate, $\hat{I}_m$ of I_m is obtained by replacing each $\gamma_k^{(m)}$ in [14.2.7] by its estimate $c_k^{(m)}$ in [14.3.2].

Pagano (1978) also demonstrates that the estimates for different periods are asymptotically uncorrelated. In other words, the joint information matrix of $\beta_1,\beta_2,\ldots,\beta_s$ is block diagonal. Consequently, the parameters for the mth season, can be estimated entirely independently of the parameters of any other season. Thus, for purposes of identification, estimation, and diagnostic checking, each season can be modelled independently of the other seasons.

When estimating the parameters in a PAR model, the orders of the AR operators can be different across the seasons. Furthermore, *subset autoregression* (McClave, 1975) can be used for constraining AR parameters to zero. For example, in season m for a monthly time series, one may wish to estimate only the AR parameters $\phi_1^{(m)}$, $\phi_2^{(m)}$ and $\phi_{12}^{(m)}$. The parameters from $\phi_3^{(m)}$ to $\phi_{11}^{(m)}$, are omitted from the model and subset autoregression is used to estimate the remaining parameters.

Model Selection using the AIC

From Section 6.3, the general formula for the AIC is defined as

$$AIC = -2\ln(ML) + 2k$$

where ML stands for the maximized value of the likelihood function and k is the number of free parameters. When using the MAICE procedure, one selects the model which gives the minimum value of the AIC.

Assuming normality, the maximized log likelihood of the AR model for season m is derived as (McLeod and Hipel, 1978)

$$\log L_m = -n\ln(\hat{\sigma}_m) + (\lambda - 1)\sum_{r=1}^{n} z_{r,m} \qquad [14.3.7]$$

The summation term on the right hand side of [14.3.7] takes into account the Jacobian of the Box-Cox transformation. An explanation of how this is done is given in Section 13.3.3 for the deseasonalized model.

The AIC formula for the mth season is

$$AIC_m = -2\log L_m + 2p_m + 4 \qquad [14.3.8]$$

where p_m is the number of AR parameters in season m. Because the mean, μ_m, and the variance of the innovations are estimated, the last term on the right hand side of [14.3.8] is included in the seasonal AIC formula.

For each combination of AR parameters, the AIC_m can be calculated using [14.3.7] and [14.3.8]. The model which yields the minimum value of the AIC_m is selected for season m. This procedure is executed for choosing the models for all of the remaining seasons. Subsequently, the AIC for the overall PAR model is

$$AIC = \sum_{m=1}^{s} AIC_m + 2 \qquad [14.3.9]$$

where the constant 2 allows for the Box-Cox parameter λ. The calculations of AIC may be repeated for several values of λ such as $\lambda = 1, 0.75, 0.5, \ldots, -1$, and the transformation yielding the minimum value of the AIC is selected.

Exhaustive Enumeration for PAR Model Selection

As mentioned earlier in Section 14.3.2, the recommended procedure for identifying the most appropriate PAR model or set of models to fit to a seasonal series is to employ the sample periodic ACF and PACF. If there is more than one promising model, the MAICE procedure can then be used to select the best one.

Another approach for determining the best AR model for each season where the maximum value of p_m is specified, would be to examine all possible regressions for that season. An appropriate criterion, such as the AIC, could be invoked for choosing the most desirable model from the exhaustive set of models. This procedure could be carried out for each season and this would result in selecting the most suitable PAR model over all of the seasons. If, for the case of a monthly series, the maximum value of p_m were restricted to be 12 for each month, the AR model for the month of March, for example, may only have AR parameters, at lags 1, 2, 3 and 12 while the other parameters would be constrained to be zero.

A possible difficulty with the aforesaid procedure is the amount of computer time required for estimating the parameters for all possible regressions for each season. For a monthly model, for example, there are 4096 possible regression models for each month and 2^{144} possible orders of monthly AR models with $p_m \leq 12$, $m = 1, 2, \ldots, 12$. Fortunately, Morgan and Tatar (1972) have devised an efficient procedure for calculating the residual sum of squares for each regression. This method drastically reduces the computational effort involved when considering an exhaustive regression study.

Because the residual sum of squares can be calculated efficiently for each regression (Morgan and Tatar, 1972), the AIC can be employed for model discrimination. In particular the residual sum of squares is used in [14.3.6] to estimate σ_m^2 and then the value of the AIC in [14.3.8] can be calculated without having to estimate the AR parameters. By selecting the model with the minimum value of the AIC, this insures that the number of model parameters are kept to a minimum and also the PAR model provides a good statistical fit to the data. Using these techniques, the best fitting PAR model for monthly data can usually be selected in less than one minute of computer time.

Subsequent, to identifying the most desirable model for season m according to the exhaustive enumeration approach, the AR parameter for this model can be estimated using subset autoregression. This procedure is repeated for each of the seasons. The value of the AIC for the overall PAR model can then be determined using [14.3.9].

A possible drawback of this exhaustive enumeration approach is that models may be identified that cannot be justified from a physical viewpoint. For instance, is it reasonable in the month of July for an average monthly riverflow series to have AR parameters for lags 2, 5 and 8? On the other hand if there were AR parameters identified for lags 1, 2 and 12, this could be justifiable from a hydrological understanding of the physical phenomenon. Applications of the exhaustive enumeration approach to average monthly riverflow time series are presented by McLeod and Hipel (1978).

14.3.4 Checking PAR Models

The adequacy of a fitted model can be ascertained by examining the properties of the residuals for each season. In particular, the residuals should be uncorrelated, normally distributed and homoscedastic.

To ascertain if the residuals are white, one must estimate the *periodic RACF* (residual autocorrelation function). For season m, the RACF at lag k is estimated using

$$r_k^{(m)}(\hat{a}_{r,m}) = \frac{\frac{1}{n}\sum_{r=1}^{n}\hat{a}_{r,m}\hat{a}_{r,m-k}}{\hat{\sigma}_m\hat{\sigma}_{m-k}}, \quad k = 1,2,\cdots \qquad [14.3.10]$$

Note that is necessary to divide by $\hat{\sigma}_m\hat{\sigma}_{m-k}$ rather than $\hat{\sigma}_m^2$ since in general $\hat{\sigma}_m \neq \hat{\sigma}_{m-k}$. Use of the incorrect divisor, $\hat{\sigma}_m^2$, could result in correlation values greater than 1.

For each season, one can plot $r_k^{(m)}(\hat{a}_{r,m})$ up to about lag $\frac{n}{4}$. Because $r_k^{(m)}(\hat{a}_{r,m})$ is asymptotically distributed as NID(0, $\frac{1}{n}$), one can also draw the 95% confidence interval for each season. If the seasonal residuals are white, they should fall within the 95% confidence limits. Nonwhiteness indicate that additional AR parameters are needed in season m or perhaps another class of models should be considered.

As a single statistic for an overall test for whiteness of the residuals, one can use the Portmanteau statistic for season m given by

$$Q'^{(m)}_L = n \sum_{k=1}^{L} (r_k^{(m)})^2(\hat{a}_{r,m}) \qquad [14.3.11]$$

This statistic is χ^2 distributed with $L - p_m$ degrees of freedom (Box and Jenkins, 1976; Box and Pierce, 1970). A significantly large value of $Q_L^{(m)}$ indicates inadequacy of the model for season m. Hence, one can reject the null hypothesis that the data in season m are white if the calculated value of $Q_L^{(m)}$ in [14.3.11] is larger than the tabulated χ^2 value at a specified significance level. One can choose L to be large enough to cover lags at which correlation could be expected to occur. For example, for monthly data, one may wish to set $L = 12$ if sufficient data are available.

As shown by McLeod (1993), a modified Portmanteau test statistic improves the small sample properties. In particular, the following exact result holds for the periodic correlations fo white noise

$$Var(r_k^{(m)}(a_{r,m})) = \frac{n - \frac{k}{s}}{n(n+2)}, \text{ if } k \equiv 0 \text{ mods}$$

$$= \frac{n - \left[\frac{k - m + s}{s}\right]}{n^2}, \text{ otherwise} \qquad [14.3.12]$$

where $[\cdot]$ denotes the integer part and $r_k^{(m)}(a_{r,m})$ is defined in [14.3.10] by replacing the residual, $\hat{a}_{r,m}$, by the theoretical innovation, $a_{r,m}$. The modified Portmanteau statistic is then defined as

$$Q_L''^{(m)} = \sum_{k=1}^{L} \frac{(r^{(m)})^2(\hat{a}_{r,m})}{\sqrt{Var(r_k^{(m)}(a_{r,m}))}} \qquad [14.3.13]$$

which is χ^2 distributed with $L - p_m$ degrees of freedom. The modified statistic in [14.3.13] reduces to that proposed for a nonseasonal ARMA model in [7.3.5] by Davies et al. (1977) and Ljung and Box (1978). One can demonstrate that

$$E\{Q_L''^{(m)}\} = L - p_m \qquad [14.3.14]$$

and

$$E\{Q_L'^{(m)}\} = n \sum_{k=1}^{L} Var(r_k^{(m)}(a_{r,m})) - p_m \qquad [14.3.15]$$

Across seasons the Portmanteau test statistics are asymptotically independent for $m = 1,2,\ldots,s$. Consequently, for the case of the Portmanteau test statistic in [14.3.13] an overall check to test if the residuals across all the seasons are white is given by

$$Q_L'' = \sum_{m=1}^{s} Q_L^{(m)} \qquad [14.3.16]$$

where Q_L'' is χ^2 distributed on $\sum_{m=1}^{s} (L - p_m)$ degrees of freedom. The lag L used in [14.3.13] could be chosen to be different across the seasons but in most applications it is reasonable to use the same value of L for all seasons. One can also employ $Q'^{(m)}_L$ in place of $Q''^{(m)}_L$ in [14.3.16] to obtain Q'_L.

One can use the tests for normality and homoscedasticity presented in Sections 7.4 and 7.5, respectively, to check that these assumptions are satisfied for the residuals in each season. These tests could also be used to ensure the assumptions hold across all of the seasons. Heteroscedasticity and/or non-normality, can often be corrected using the Box-Cox transformation in [13.2.1].

14.4 PAR MODELLING APPLICATION

The model construction techniques of Section 14.3 are employed for determining the most appropriate PAR model to fit to the average monthly flows of the Saugeen River at Walkerton, Ontario, Canada, which are available from Environment Canada (1977) from January, 1916 to December, 1976. From the sinusoidal structure contained in the graph of the last ten years of the average monthly Saugeen riverflows shown in Figure VI.1, one can see that the observations are highly seasonal. As emphasized in Section 14.3.2 and Chapter 15, the recommended approach for identifying the AR parameters required in each season for the PAR model is to use the sample periodic ACF and PACF. Because it is known a priori that most average monthly riverflow series require a natural logarithmic transformation to avoid problems with the residuals of the fitted model, the logarithmic Saugeen flows are used right at the start of the identification stage.

Figure 14.4.1, displays the graph of the periodic ACF against lag k for the logarithmic monthly Saugeen riverflows. Notice that each period or month possesses an ACF which is plotted vertically. The two lines above a given period show the 95% confidence interval. To keep the graph simple, the zero line, which falls midway between the confidence interval, is not given. Opposite a particular lag, the estimated value of the ACF for a given period is plotted horizontally. If the line cuts the left or right line for the confidence interval, the value of the sample ACF is significantly different from zero. Notice in Figure 14.4.1 that the estimated periodic ACF at lag 1 is significantly different from zero for all periods or months except for March (period 3) where the value just touches the 95% confidence limits. Because flows in one month are usually correlated with flows in the previous month, this behaviour would be expected. In addition, for some of the months such as January, October, November and December, which are indicated by periods 1, 10, 11, and 12, respectively, it appears that the ACF may be attenuating.

To identify more clearly the order of the AR model in each season, one must examine the sample periodic PACF plotted in Figure 14.4.2. Notice that the sample PACF for each period or season $m = 1,2,\ldots,12$, is plotted vertically along with the 95% confidence interval. There are significantly large values of the sample PACF at lag 1 across all 12 of the months, although in period 3 or March, the sample PACF is only just touching the 95% confidence interval. Furthermore, for all the months the sample PACF truncates and is not significantly different from zero after lag 1. Therefore, the identification plots indicate that for all months except possibly March, one should use an AR model of order 1 or a Markov model.

The parameters in the PAR(1,1,0,1,1,1,1,1,1,1,1,1) model are estimated using the periodic Yule-Walker equations in [4.2.12] for each season. The fitted model satisfies the tests for whiteness, heteroscedasticity and normality described in Section 14.3.4. For example, when the sample periodic RACF is plotted, the assumption of whiteness for the values of the RACF for each of the months is reasonably well satisfied. In particular, Figure 14.4.3 shows a graph of the periodic RACF calculated using [14.3.10] for the calibrated PAR model fitted to the logarithmic average monthly Saugeen riverflows. Notice that for all of the months, or periods, at most one value falls outside the 95% confidence limits which are calculated assuming that the RACF values are asymptotically NID$(0,\frac{1}{n})$. Moreover, at the crucial first few lags as well as lag 12, none of the RACF values are significantly different from zero for any of the seasons.

For both the periodic ACF and PACF graphs shown in Figures 14.4.1 and 14.4.2, respectively, at the fourth month or period there is a significantly large negative correlation at lag one. One way to interpret this behaviour from a physical viewpoint is that when spring flows in March cause large March flows due to the snowmelt runoff, the April flows tend to be substantially smaller.

Table 14.4.1 provides the parameter estimate and SE (standard error) for the AR parameter at lag one for each of the twelve seasons or periods for the PAR model fitted to the logarithmic average monthly Saugeen flows. One can see that the estimates reflect what is found in the periodic ACF and PACF plots in Figures 14.4.1 and 14.4.2, respectively. In particular, the AR parameter estimate for April is negative, as is also the case for the values of both the sample periodic ACF and PACF at lag one in period four.

Table 14.4.1. Parameter estimates and SE's for the PAR model having one AR parameter in each season, except for March, that is fitted to the logarithmic average monthly Saugeen riverflows.

Seasons or Periods	AR Parameter Estimates	SE's
1	0.6472	0.1037
2	0.4977	0.0886
3	0	0
4	-0.3124	0.0916
5	0.5300	0.1057
6	0.6091	0.0943
7	0.7087	0.1169
8	0.4228	0.0730
9	0.7039	0.1150
10	1.0598	0.1238
11	0.7699	0.0828
12	0.5901	0.1015

An advantage of employing the PAR model is that it can capture the type of varying seasonal correlation structure just described. Because of this, one would expect that the PAR model would more accurately and realistically describe the behaviour of the monthly Saugeen riverflows than competing types of seasonal models. This fact is confirmed by comparing the calculated value of the AIC in [14.3.9] for the Saugeen PAR model to those computed for the best SARIMA and deseasonalized models fitted to the average monthly Saugeen riverflow series in

Sections 12.4.4 and 13.4.2, respectively. The AIC value of 3357.82 for the PAR model is substantially less than those calculated for the other two models. Consequently, the PAR model is recommended over the SARIMA and deseasonalized models for fitting to the monthly Saugeen riverflows. Likewise, the forecasting experiments in Chapter 15, demonstrate that PAR models forecast average monthly riverflow series more accurately than its competitors and, therefore, is a better model to use with this type of seasonal data.

14.5 PARSIMONIOUS PERIODIC AUTOREGRESSIVE (PPAR) MODELS

14.5.1 Introduction

The PAR models described in the previous sections of this chapter attempt to preserve the seasonally-varying autocorrelation structure of a time series by fitting a separate AR model to each season of the year. However, one could reasonably question the necessity of going to the extreme of having a different model for each and every season. To decrease the number of model parameters required in a PAR model, one could combine individual AR models for various seasons in order to obtain a single model for all seasons in a given group. After grouping, the parameters of the more *parsimonious PAR* or *PPAR models* are estimated and diagnostically checked, and the PAR and PPAR models compared.

The approach for developing a PPAR model described in this section was originally presented by Thompstone et al. (1985a) and also Thompstone (1983). As an alternative procedure for reducing the number of parameters in PAR or PARMA models, Salas et al. (1980) propose a Fourier series approach. Recall that a Fourier series procedure is presented in Section 13.3.3 for reducing the number of deseasonalization parameters needed in the deseasonalized models of Chapter 13.

In the next subsection, the PPAR model is formally defined. Following this, flexible model construction techniques are given. In Section 14.6, all of the seasonal models of Part VI are compared by fitting them to six hydrological time series.

14.5.2 Definition of PPAR Models

As is also the case for the PAR model in [14.2.3], let the number of years and seasons be n and s, respectively, and let a transformed observation be given by $z_{r,m}^{(\lambda)}$, $r = 1,2,\ldots,n$, and $m = 1,2,\ldots,s$. Assuming the s seasons are grouped into G groups of one or more seasons with similar AR characteristics, the *parsimonious periodic autoregressive model (PPAR)* written as $(p_1,p_2,\ldots,p_G)$ may be defined in a manner analogous to the PAR model in [14.2.3] as

$$\phi^{(g)}(B)(z_{r,m}^{(\lambda)} - \mu_m) = a_{r,m} \qquad [14.5.1]$$

where $\phi^{(g)}(B) = 1 - \phi_1^{(g)}B - \phi_2^{(g)}B^2 - \cdots - \phi_{p_g}^{(g)}B^{p_g}$, is the AR operator of order p_g for group g, μ_m is the mean for season m, and $a_{r,m} \approx NID(0,\sigma_g^2)$. Notice from equation [14.5.1] that within a given group each seasonal mean is preserved by the parameter μ_m. However, for the observations in the seasons contained in the gth group, the AR parameters and the variance of the residuals are assumed to be the same.

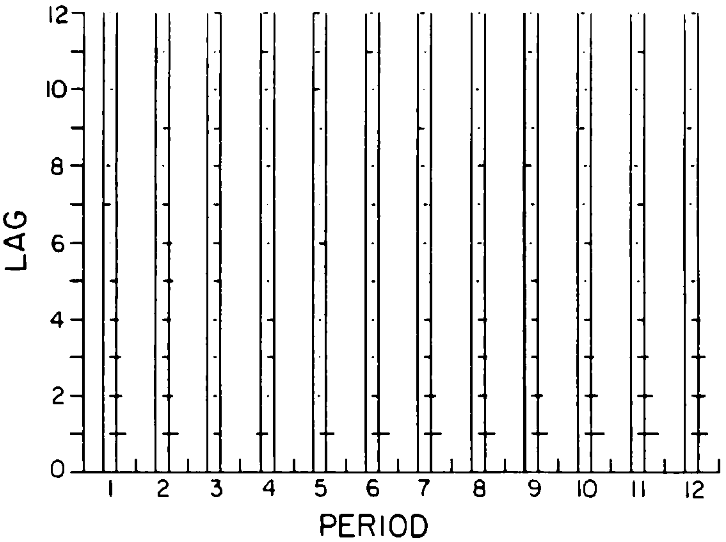


Figure 14.4.1. Sample periodic ACF for the logarithmic average monthly flows of the Saugeen River from January, 1916, until December, 1976, at Walkerton, Ontario, Canada.

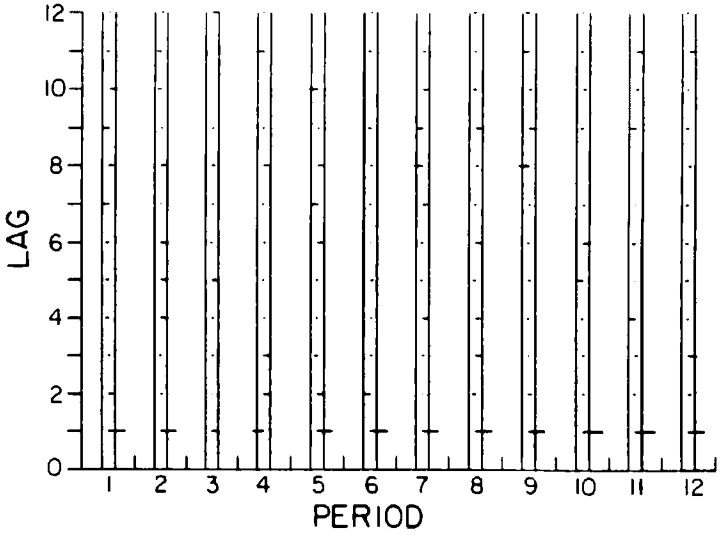


Figure 14.4.2. Sample periodic PACF for the logarithmic average monthly flows of the Saugeen River from January, 1916, until December, 1976, at Walkerton, Ontario, Canada.

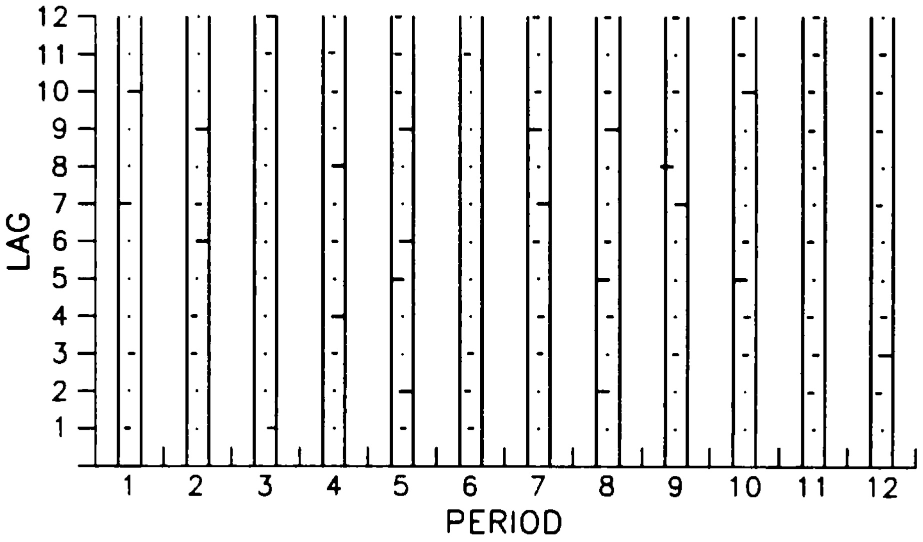


Figure 14.4.3. Periodic RACF for the PAR model having one AR parameter for each season, except for March, fitted to the average monthly riverflows of the Saugeen River from January, 1916, until December, 1976, at Walkerton, Ontario, Canada.

14.5.3 Constructing PPAR Models

In order to identify an appropriate grouping of seasons, the approach examined herein involves first fitting PAR models to the time series in question as described in Section 14.3. One then attempts to find seasons for which the AR models are "compatible". The equation of season m_2 is said to be compatible with that of season m_1, if the residuals obtained when the equation fit to season m_2 is applied to season m_1 are not significantly different from the residuals obtained from the equation fit to season m_1. In order to test formally for compatibility, define $a_T(m_1,m_2)$ to be the residuals obtained when the model fit to season m_2 is applied to season m_1 using [14.2.3] with initial values set to zero. These residuals can be used to estimate $\sigma^2(m_1,m_2)$, the residual variance when the model for season m_2 is applied to season m_1.

Consider the null hypothesis

$$H_o:\sigma^2(m_1,m_2)=\sigma^2(m_1,m_1)$$

Assuming that $(a_{R,M}(m_1,m_2),a_{R,M}(m_1,m_1))$ are jointly normally distributed with mean zero and are independent for successive values, a test developed by Pitman (1939) can be used to test this null hypothesis. For a review of how to carry out a hypothesis test, the reader can refer to Section 23.2. Let

$$S_{R,M} = a_{R,M}(m_1,m_2) + a_{R,M}(m_1,m_1) \qquad [14.5.2]$$

and

$$D_{R,M} = a_{R,M}(m_1,m_2) - a_{R,M}(m_1,m_1) \qquad [14.5.3]$$

Pitman's test is then equivalent to testing if the correlation, ρ, between $S_{R,M}$ and $D_{R,M}$ is significantly different from zero. Thus, provided $n > 25$, H_0 would be accepted at the 5% level of significance if $|\rho| < 1.96/\sqrt{n}$.

In practice, the residuals may not satisfy exactly the assumptions of a joint normal distribution with mean zero and independence for successive values of the residuals. However, these assumptions are probably a sensible first approximation. The assumption of independence seems reasonable because, with annual periodicity, the residuals are chronologically one year apart. Furthermore, the mean of zero is assured for the case of $a_{R,M}(m_1,m_1)$ due to the method of fitting the model. Pitman's test has often been used for testing the equality of variances of paired samples (Snedecor and Chochran, 1980, p. 190). It was pointed out in Lehmann (1959, p. 208, problem 33) that in this situation the test is unbiased and uniformly most powerful.

The above definition of equation compatibility can be extended to mutual compatibility. In particular, equations for seasons m_1 and m_2 are mutually compatible, if, at a given level of significance, one would accept the following two hypotheses:

$$\sigma^2(m_2,m_1) = \sigma^2(m_1,m_1)$$

$$\sigma^2(m_1,m_2) = \sigma^2(m_2,m_2)$$

Thus, the criteria adopted herein for identifying seasons in the same group is that each pair of seasons in the group must be mutually compatible at a given level of significance and have the same order of AR model. In addition, seasons are not grouped together unless they are chronologically adjacent. Once the groups have been identified, the parameters are estimated using maximum likelihood estimation. Specifically, multiple linear regression can be used to estimate the AR parameters for each group of seasons, where the seasonal means are estimated using [14.3.5] and the estimated variance of the residuals for each season is calculated using the estimated residuals contained in the group of seasons. Diagnostic checking involves first calculating the residuals from [14.5.1] by setting initial values to zero, and then examining the seasonal RACF and related Portmanteau test statistics plus tests for normality and homoscedasticity.

For season m in a PAR model, the maximized log likelihood is presented in [14.3.7]. When considering a PPAR model, the maximized log likelihood for the gth group is

$$\log L_g = -n_g \ln(\hat{\sigma}_g) + (\lambda - 1) \sum_{z_{r,m} \in \text{ group } g} z_{r,m} \qquad [14.5.4]$$

where n_g is the product of the number of seasons in group g and the number of years of data, n. Notice that the summation term on the right hand side of [14.5.4] is for all data points contained in the seasons in the gth group. The value of the maximized log likelihood can be obtained by summing [14.5.4] across all the seasons to obtain

$$L_{PPAR} = \sum_{g=1}^{G} \log L_g \qquad [14.5.5]$$

Likewise, for a PAR model the value of the maximized log likelihood across all s seasons is

$$L_{PAR} = \sum_{m=1}^{s} \log L_m \qquad [14.5.6]$$

where $\log L_m$ is defined in [14.3.7].

As was done for the PAR model, one can derive the AIC for a PPAR model for each group of seasons and also the overall model. In particular, for the gth group of seasons the AIC formula is

$$AIC_g = -2\log L_g + 2p_g + 2 + 2\,(\textit{number of means}) \qquad [14.5.7]$$

where p_g is the number of AR parameters in the gth group seasons. The other parameters are the variance of the residuals and the number of means in the gth group of seasons. The AIC for the overall PPAR model is determined as

$$AIC_G = \sum_{g=1}^{G} AIC_g + 2 \qquad [14.5.8]$$

where the constant 2 allows for the Box-Cox parameter λ. The overall AIC formula for the PAR model is presented in [14.3.9].

When both PAR and PPAR models are fitted to a given series, the log-likelihood ratio (Rao, 1973, p. 448) can be used to test the null hypothesis that there is no significant difference in the residuals of the two models. It may be expressed as

$$\bar{R} = -2[L_{PPAR} - L_{PAR}] \qquad [14.5.9]$$

and, assuming the null hypothesis is true, $\bar{R}$ follows a chi-squared distribution with the number of degrees of freedom equal to the difference in the number of free parameters in the PAR and PPAR models, respectively (i.e., the difference in the number of AR parameters and residual variances).

14.6 APPLICATIONS OF SEASONAL MODELS

All of the seasonal models presented in Part VI are fitted to three average monthly riverflow series and three average quarter-monthly riverflow time series and the resulting models are compared using the AIC. More specifically, the seasonal models fitted to the series are the SARIMA, deseasonalized, PAR and PPAR models defined in Sections 12.2, 13.2, 14.2.2 and 14.5.2, respectively. Grouping of seasons within the PPAR models is performed using three levels of significance in the Pitman test presented in Section 14.5.3, namely 50%, 20% and 5%. In general, as the level of significance decreases, fewer seasons are considered to have "incompatible" models and thus there is more grouping, or in other words, a smaller number of groups. A Box-Cox transformation with λ = 0 is used in all cases and, hence, the data are transformed by taking their natural logarithms. The above mentioned models are labelled as SARIMA, DES, PAR, PPAR/50, PPAR/20 and PPAR/05, respectively, in the upcoming tables. The results of this study were originally presented by Thompstone (1983, Section 3.5).

The six example hydrological time series consist of:

(1) inflows to reservoirs of the hydroelectric system operated by Alcan Smelters and Chemicals Ltd. in the Province of Quebec, Canada (Thompstone et al., 1980);

(2) flows of the Saugeen River measured at Walkerton, Ontario, Canada (Environment Canada, 1977);

(3) flows of the Rio Grande measured at Furnas, Minas Gerais, Brazil (supplied by Mr. Paulo Roberto de Holanda Sales at Eletrobras, (national electrical company of Brazil)).

The monthly series are comprised of Alcan system inflows from 1943 to 1979, Saugeen River flows, 1919-76, and Rio Grande flows, 1931-75; the quarter-monthly flows consist of Alcan system inflows, 1943-79, Saugeen riverflows, 1915-76, and Rio Grande flows, 1931-72. Note that the quarter-monthly data consists of flows in m^3/s averaged from the 1st to the 7th, from the 8th to the 15th, from the 16th to the 22nd, and from the 23rd to the end of the month, which constitute periods of approximately one week each.

For all six series, the order of the AR operator in a PAR or PPAR model for a season or group of seasons, respectively, is usually one while the highest order is three. Very few of the AR models for an individual season or group of seasons are white noise.

Table 14.6.1 summarizes the orders of the AR models contained within the PAR models fitted to the six series. Because there are 48 and 12 seasons for the quarter-monthly and monthly data, respectively, the number of AR models used in each quarter-monthly PAR model must equal 48 whereas the total for each PAR model is 12. For the case of the PAR model for the average monthly Saugeen riverflows, the order of the AR operator is one for 11 of the 12 months. As explained in Section 14.4, the month of March is white noise. Finally, the only other series which has white noise components in the PAR model is the Alcan system for monthly riverflows that contains four such months.

In order to illustrate the degree of grouping associated with various Pitman test significance levels, Table 14.6.2 shows the number of groups associated with the PAR, PPAR/50, PPAR/20 and PPAR/05 models for each series. For the case of quarter-monthly series, the highest degree of grouping is with the PPAR/05 model of Rio Grande flows: the 48 seasons are divided into 16 groups. The highest degree of grouping of monthly series is with the PPAR/05 model of the Saugeen riverflows: the 12 months are divided into 5 groups. Note that, in the case of the Alcan system monthly inflow series, no grouping is identified, even when using the 50% significance level.

Table 14.6.3 shows how all six of the seasonal models are ranked according to the AIC for each of the series. The model having the lowest AIC value is ranked first whereas the one with the highest value is ranked as 6. When fitting the deseasonalized model, the logarithmic series is fully deseasonalized using [13.2.3]. Although it isn't done in this study, one could reduce the number of deseasonalization parameters by implementing the Fourier series approach described in Section 13.3.3.

As shown in Table 14.6.3, the AIC always selects a PPAR model as the most desirable model. The only exception is the PAR model for the Alcan system for which no PPAR model is identified. As would be expected from the basic design of the SARIMA model, in all six cases the SARIMA model is the least desirable model, according to the AIC. This is because the

Table 14.6.1. Number of periods having a given order of an AR model for the PAR models fitted to six series.

Order of AR Model	Quarter-monthly Series			Monthly Series		
	Alcan System	Rio Grande	Saugeen	Alcan System	Rio Grande	Saugeen
0	0	0	0	4	0	1
1	36	42	43	6	9	11
2	10	6	5	2	2	0
3	2	0	0	0	1	0

Table 14.6.2. Number of groups in the PAR and PPAR models.

Model	Quarter-monthly Series			Monthly Series		
	Alcan System	Rio Grande	Saugeen	Alcan System	Rio Grande	Saugeen
PAR	48	48	48	12	12	12
PPAR/50	40	38	40	12	10	8
PPAR/20	29	27	23	12	8	7
PPAR/05	24	16	18	12	6	5

Table 14.6.3 Ranking of the seasonal models fitted to the six series according to the AIC.

Model	Quarter-monthly Series			Monthly Series		
	Alcan System	Rio Grande	Saugeen	Alcan System	Rio Grande	Saugeen
SARIMA	6	6	6	3	6	6
DES	5	5	5	2	5	5
PAR	4	3	3	1	4	4
PPAR/50	3	1	2	-	2	3
PPAR/20	1	2	1	-	1	1
PPAR/05	2	4	4	-	3	2

SARIMA model is not designed for describing stationarity within each season as well as a seasonally varying correlation structure. Because the deseasonalized model of Chapter 13 can account for a separate seasonal mean and variance within each season, the AIC results of Table 14.6.3 indicate that the deseasonalized model always performs better than the SARIMA in all six applications. Moreover, due to the fact that a periodic model can handle seasonally varying correlation, periodic models always do better than both deseasonalized and SARIMA models. Finally, forecasting experiments are carried out in Section 15.4.4 to compare the forecasting capabilities of the models listed in Table 14.6.3.

The log-likelihood ratio test in [14.5.9] can be used to ascertain if the residuals of the fitted PPAR and PAR models differ significantly from each other. In the five cases for which PPAR models are identified (see Tables 14.6.2 or 14.6.3), residuals of none of the PPAR models are significantly different from those of the PAR model at the 5% level of significance. This reinforces the conclusion that even though PPAR models have fewer parameters and, hence, also the seasonal models, they still describe the data as well as the regular PAR model.

As noted in Section 14.5.1, another approach for reducing the number of AR parameters required in a PAR model is to use a Fourier series approach (Salas et al., 1980). However, this procedure assumes that a smooth sinusoidal type of curve is fitted to the AR parameters across the seasons or periods and, hence, also the sample periodic ACF. The question arises as to whether this assumption is reasonable. To investigate this, consider Figure 14.6.1, which shows a graph of the periodic ACF at lag one of the natural logarithms of the quarter-monthly flows of the Saugeen River. As can be seen, it would be impossible to fit a smooth cyclic curve through this plot. Notice, for example, the manner in which the first order correlation drops significantly downwards in the spring season (i.e., about the end of March in the 12th quarter-monthly period). Fortunately, both the PPAR and the PAR models are designed for modelling the type of behaviour exhibited in Figure 14.6.1. The approach to fitting PAR and PPAR models is sufficiently general to be applicable to series with or without a cyclic pattern in the seasonal correlations and AR parameters. In a similar fashion, one can see that it would be difficult to fit a Fourier series curve through the AR parameter estimates in Table 14.4.1 calculated for the PAR model fitted to the logarithmic average monthly Saugeen riverflows.

14.7 CONSTRUCTING PARMA MODELS

PARMA models can be fitted to seasonal series by following the identification, estimation and diagnostic check stages of model construction. Because model building procedures are highly developed for use with PAR models, this class of periodic models is focussed upon in this chapter. Nonetheless, there are now some good construction techniques available for fitting PARMA models to seasonal data sets. As is also the case for the PAR model, the ARMA model for each season of the year can be identified separately. The main area where further research is required for PARMA model building is the development of a maximum likelihood estimation technique which is computationally efficient. To obtain efficient estimates for a PARMA model, all parameters must be estimated simultaneously, including the innovation variances, and, moreover, it is necessary to use a nonlinear optimization technique since the likelihood function is nonlinear. Each evaluation of the likelihood function involves very lengthy computations when $s \geq 12$.

The sample periodic ACF and PACF described in Section 14.3.2 can be employed for identifying the orders of the AR and MA operators for the PARMA model in [14.2.15] to fit to each of the seasons in a given seasonal time series. If a pure MA model of order q_m is required, the sample periodic ACF for season m will not be significantly different from zero after lag q_m and the sample periodic PACF will die off. When a pure AR model of order p_m is needed to model season m, the sample periodic ACF attenuates while the sample periodic PACF is not significantly different from zero after lag p_m. When both AR and MA parameters should be included in the ARMA model to fit to the mth season, both the sample periodic ACF and PACF attenuate.

Assuming normality, Vecchia (1985a,b) developed a technique for obtaining MLE's of the parameters in a PARMA model. The approach that Vecchia (1985a,b) uses to write the likelihood function is the same as the one of Newbold (1974) for the univariate case and Hillmer and Tiao (1979) for the multivariate ARMA models presented in Chapter 20. Additionally, he proved that PARMA models and multivariate ARMA models are equivalent. From a computational point of view, his algorithm seems to be feasible for use in practical applications when the number of seasons is small (i.e., less than about 4 seasons per year). To overcome computational

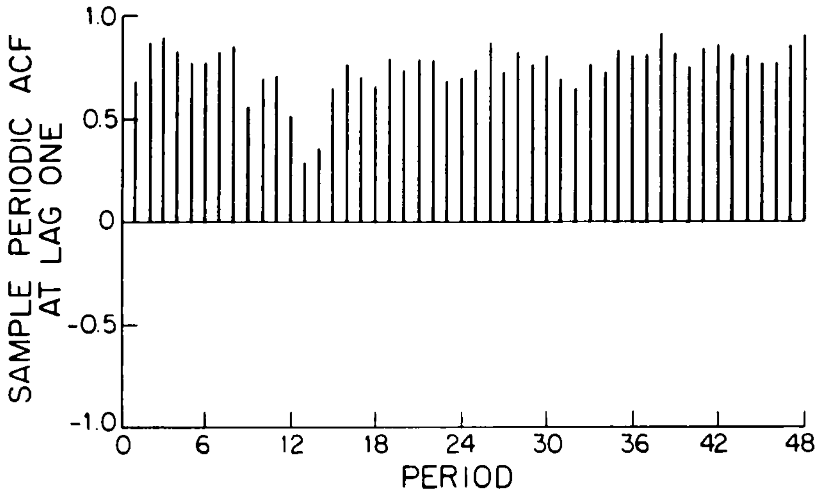


Figure 14.6.1. Periodic ACF at lag one of the logarithmic quarter-monthly riverflows of the Saugeen River at Walkerton, Ontario, Canada, from 1915 to 1976.

difficulties, Jimenez et al. (1989) propose a maximum likelihood parameter estimation technique which is implemented within a Kalman filtering framework. Finally, Li and Hui (1988) provide an algorithm for the exact likelihood of PARMA models.

As explained in Section 14.3.3, the Yule-Walker equations can be used as a moment estimation approach for obtaining efficient parameter estimates for the parameters of a PAR model. However, one should be cautious when using moment estimators with PARMA models, since the parameters estimates may not be efficient. Nonetheless, some research on moment estimation of PARMA model parameters has been completed. For example, Salas et al. (1982) derived Yule-Walker equations for PARMA models and showed how moment estimates can be calculated for PARMA models in which $p_m \geq 0$ and $q_m = 1$ in season m. Besides discussing moment estimation, Salas and Obeysekera (1992) described model identification and testing of model adequacy of PARMA models. Moreover, these authors proved a physical basis for PARMA models. In particular, based upon a conceptual-physical representation of a natural watershed, in which all inputs, storages, outputs and parameters are assumed to be periodic and the system is a linear reservoir, they demonstrated that the periodic groundwater storage and streamflow processes belong to the class of PARMA processes. Section 3.6 describes this kind of physical relationships for the case of nonseasonal ARMA models. Further results on how PARMA models can be used in physically-based modelling are provided by Claps et al. (1993).

When testing the adequacy of a calibrated PARMA model, one can use similar procedures to those suggested for PAR models in Section 14.3.4. The sample periodic RACF and related Portmanteau statistics can be employed to ascertain if the residuals are white. Other tests related to those presented in Chapter 7 for nonseasonal ARMA models can be used for testing if the normality and homoscedasticity assumptions are valid. Non-normality and/or heteroscedasticity can often be rectified by incorporating an appropriate Box-Cox transformation from [13.2.1].

In related research to PARMA modelling, Vecchia et al. (1983) investigated what happens when one aggregates across the seasons. Specifically, the aggregated time series resulting from summing over the seasons of a seasonal time series, which is assumed to be either AR(1) or ARMA(1,1) in each season, is shown to follow an ARMA(1,1) model at the annual level. Moreover, when the seasonal data and the model for each season are used rather than the annual data and the associated annual model, significant gain in parameter efficiency can be achieved. This, of course, further justifies the use of PAR and PARMA modelling in water resources and indicates that aggregation is preferable to disaggregation. A discussion of disaggregation and the controversy surrounding it is given in Section 20.5.2.

For most of the PARMA model construction techniques discussed thus far, it is assumed that the data or, equivalently, the model residuals are normally distributed. Fernandez and Salas (1986) studied PAR models having a Gamma marginal distribution. This Gamma or other kinds of distributional assumption could also be used with PARMA models. However, a substantial amount of theoretical research and development of flexible model building techniques are needed before these and other related models can be used in practice. Lewis (1985) and authors referenced therein, discuss non-Gaussian distributed innovations for use in nonseasonal and multivariate modelling. In Section 20.5.3, the employment of non-Gaussian marginal distributions in multivariate modelling is outlined.

14.8 SIMULATING AND FORECASTING WITH PERIODIC MODELS

14.8.1 Introduction

Subsequent to fitting a PAR or PARMA model to a seasonal time series, the calibrated model can be used for applications such as forecasting and simulation. In the next chapter, it is explained how minimum mean squared error forecasts from periodic models, as well as other kinds of seasonal models, can be calculated. Moreover, forecasting experiments with average monthly riverflow series demonstrate that PAR models forecast better than deseasonalized (Chapter 13) and SARIMA (Chapter 12) models.

In Chapter 9, two simulation procedures are presented for generating synthetic data from nonseasonal AR and ARMA models. The simulation techniques are designed so that random realizations of the underlying stochastic process are employed as starting values. Because fixed beginning values are not utilized, unwanted systematic bias is not introduced into the synthetic traces.

Because a PAR or PARMA model consists of having a separate AR or ARMA for each season of the year, simulation techniques similar to those presented in Chapter 9 for use with nonseasonal models can be employed with seasonal models. The technique labelled WASIM2, for example, in Section 9.4 exactly simulates an AR or ARMA process if the residuals are assumed to be normally distributed. Suppose, for example, one wishes to simulate using a PAR model. Let $k = \max(p_1, p_2 - 1, p_3 - 2, \ldots, p_s - (s - 1))$ where s is the number of seasons. By utilizing the covariance matrix of $(z_{1,1}, z_{1,2}, \ldots, z_{1,k})$ to generate randomly the initial values, a technique very similar to WASIM2 can be used for producing synthetic traces from a PAR model. If deemed appropriate, parameter uncertainty can also be brought into a simulation study by following the WASIM3 procedure of Section 9.7. Salas and Abdelmohsen (1993) describe initialization techniques when simulation with single-site and multisite low-order PAR and PARMA models.

As explained in Chapter 9, simulation can be used for design purposes and investigating the theoretical properties of models. In the next subsection, it is shown using simulation that PAR models can preserve statistically the critical drought statistics defined by Hall et al. (1969).

Stedinger and Taylor (1982a,b) describe the steps involved in the development and use of a stochastic streamflow model. After properly fitting a time series model to a given nonseasonal or seasonal riverflow data set, these authors stress the importance of model verification and model validation. In model verification, one should demonstrate that a model has been implemented correctly and passes diagnostic checks. With respect to model validation, one should show that simulated sequences from the calibrated model produce reservoir system performance that is consistent with or statistically indistinguishable from that obtained utilizing the historical riverflows. Accordingly, the simulation experiments carried out in Section 14.8.2 as well as Section 10.6 can be considered to be model validations.

14.8.2 Preservation of Critical Period Statistics

Introduction

Hall et al. (1969) discuss problems related to the design and operation of a reservoir when water shortages must be considered. They define the critical period as the period of time for which a given inflow series is most critical with respect to meeting water demands. Various statistics, which are closely related to the critical period, are defined and, by using simulation, Hall et al. (1969) conclude that the stochastic model they are investigating does not adequately preserve the historical critical period statistics. In a more exhaustive study, Askew et al. (1971) find that a large variety of stochastic models are not capable of retaining the critical period statistics. The purpose of the present section is to demonstrate that, for certain sample series, when the PAR and PPAR models are identified and fitted using the procedures described in this chapter, they adequately preserve the historical critical period statistics.

Critical Periodic Statistics for Water Supply

Hall et al. (1969) express the active reservoir storage as a ratio of the total volume of active storage in the reservoir to the volume of water due to the average annual inflow. The reservoir is operated to allow a *seasonal extraction* of X. It is assumed that the reservoir is full at the start and a value of X is determined which causes the reservoir storage to research zero at one point in time. The *length of the critical period* is denoted by L and is calculated as the time span from the zero storage point backward in time to the point when the reservoir was last full. The *percentage deficiency, D, for the critical period* is defined as

$$D = \frac{\sum_{CP}(\bar{V} - V_t)100}{\bar{V}L} \qquad [14.8.1]$$

where $\bar{V}$ is the average seasonal inflow volume, V_t is the seasonal inflow volume for period t, and the summation extends over the entire critical period, CP. As pointed out by Hall et al.

(1969), the aforesaid critical period statistics can be readily generalized to the case where the extraction is a function of time, the reservoir is at any level at the start of the calculations, and evaporation and other losses are considered. Note also that the critical period statistics are obviously a function of the length of the series for which they are defined. As illustrated by McMahon and Mein (1978, pp. 19-20), there may, in rare cases, be more than one critical period for a given inflow series.

Design of Simulation Experiments

In Section 14.6, PAR and PPAR models are fitted to three quarter-monthly and three-monthly time series. These same series are used in this section in split sample simulation experiments used to show that PAR and PPAR models preserve statistically the critical period statistics. These results were originally presented by Thompstone et al. (1987) and Thompstone (1983). McLeod and Hipel (1978) used simulation experiments to demonstrate that critical period statistics are preserved by PAR models but they did not use the split sample approach described herein. Recall that in Sections 9.8 and 10.6, simulation experiments are used to demonstrate that ARMA models preserve statistically the rescaled adjusted range and other statistics related to the Hurst phenomenon.

For each of the 6 seasonal series, PAR and PPAR models are identified and fit following the procedures of Sections 14.3 and 14.5.3, respectively. In each case, all but the last 20 years of available data are used to identify and fit the models. The PPAR model having the minimum value of the AIC is selected from three candidates, namely those with 50%, 20% and 5% levels of significance for the Pitman test grouping criterion described in Section 14.5.3.

As explained in Section 9.2, in order to generate synthetic sequences, it is first necessary to produce independent, normally distributed random numbers with a mean of zero and a variance of one. In the experiments described herein, an efficient and portable pseudo-random number generator, developed by Wickmann and Hill (1982), is used to produce numbers rectangularly distributed between zero and one, and these are then used in the algorithm of Box and Muller (1958) to produce the required random normal deviates. These innovations are then fed into the appropriately estimated PAR model in [14.2.3] or the PPAR model in [14.5.1] for a given series.

As pointed out in Chapter 9, an important consideration in the generation of synthetic hydrological sequences is the choice of initial values. Random realizations of the underlying stochastic process must be used to avoid introducing systematic bias into the simulation study. The approach to obtaining random realizations adopted in the original study by Thompstone et al. (1987) and Thompstone (1983) is to, in a preliminary study, set the required initial values to their expected values and then generate a full 40 years of synthetic data. The last few values of these 40 years of synthetic data provide the required initial values for the main simulation study.

For a given sample time series, the simulation experiment is conducted as follows. First, the remaining 20 years of the historical sample not used in model construction are used to calculate what are referred to as the historical critical period statistics. These are denoted as X(his), L(his) and D(his) for the historical extraction rate, historical length of the critical period, and historical deficiency, respectively. An active reservoir storage equal to the average volume of annual inflow is used. Next, 1,000 synthetic seasonal sequences of 40 years each are generated, and the first 20 years of each sample are dropped to provide 1,000 effectively independent sequences equal in length to the series used to calculate the historical critical period statistics.

It is important to note that almost all previous research concerning the preservation of statistics in synthetic hydrological sequence generation has not used the split-sample approach employed herein. In previous research, the same sample series was employed both to construct the model(s) being evaluated and to estimate the statistic(s) whose preservation is being studied. One would generally expect the split-sample design of the current research to be a more rigorous validation of the models under investigation.

In order to test if a given model preserves the critical period statistics, the P-values defined below are estimated:

$$P_X = Prob\{X(syn) < X(his)\} \qquad [14.8.2]$$

$$P_L = Prob\{L(syn) > L(his)\} \qquad [14.8.3]$$

$$P_D = Prob\{D(syn) > D(his)\} \qquad [14.8.4]$$

where *Prob* denotes probability, $X(syn)$ is the extraction rate in the synthetic series, $L(syn)$ is the length of the critical period in the synthetic series, $D(syn)$ is the percentage deficiency in the synthetic series, and other terms are as defined earlier.

The P-values are estimated separately for each series with the active reservoir storage equal to the volume of the average annual inflow for the 20-year historical sample not used to calibrate the models. This is done by counting the number of times the inequalities in [14.8.2] to [14.8.4] hold in each simulation run and dividing by 1,000. The P-values, as defined above, represent the probability of a critical period statistic in the synthetic sequence being more extreme than in the historical sequence. Thus a P-value of 0.05 indicates that there is only a 5% chance that the synthetic series will have a critical period statistic more extreme than the historical. Of course, this would happen 5% of the time even if the historical sequence were in fact generated by the corresponding fitted stochastic model. Nevertheless, P-values less than 5% do suggest possible model inadequacy, and hence, P-values can be used for diagnostic checking.

In Section 10.6.4, a χ^2 test is employed to ascertain, in an overall sense, if the Hurst statistics are preserved statistically by ARMA models fitted to 23 annual geophysical time series. In particular, when considering k time series for a given statistic, it can be shown (Fisher, 1970, p. 99)

$$-2\sum_{i=1}^{k}\ln(P_i) \approx \chi_{2k}^2 \qquad [14.8.5]$$

where P_i can be the probability as defined in Equations [14.8.2] to [14.8.4] for the ith time series.

The Results of the Simulation Experiments

The results of the simulation experiments are summarized first for the PAR models, and then for the PPAR models. Table 14.8.1 shows the P-values for PAR models for the three critical period statistics and for the six example series, while Table 14.8.2 contains the chi-squared values calculated using [14.8.5] for the three critical period statistics with the series grouped according to their seasonal lengths. For a one-sided significance test, the chi-squared values with six degrees of freedom at the 5% and 1% significance levels are 12.592 and 16.812, respectively. For the monthly series, the critical statistics are preserved in each case at the 5% level, as can be seen in Table 14.8.1, and on an overall basis, also at the 5% level, as shown in Table 14.8.2.

For the quarter-monthly series, evidence of preservation of the critical period statistics by PAR models is not quite as strong. The extraction rate for the Saugeen series and the length of the critical period for the Alcan system inflows are preserved at the 1% level, but not at the 5% level. In all other cases, the statistics are preserved at the 5% level. According to the overall chi-squared test, the length of the critical period and deficiency percentage are preserved at the 5% level, but the extraction rate is preserved at only the 1% level.

Table 14.8.1. P-values for the PAR models.

Riverflow Series		Statistics		
		Extraction	Length of CP	Deficiency
Quarter-Monthly	Alcan System	0.233	0.040	0.730
	Rio Grande	0.350	0.267	0.654
	Saugeen	0.017	0.267	0.257
Monthly	Alcan System	0.423	0.118	0.774
	Rio Grande	0.521	0.090	0.825
	Saugeen	0.083	0.370	0.292

Table 14.8.2. Chi-squared values for the PAR models.

Seasonal Lengths	Statistics		
	Extraction	Length of CP	Deficiency
Quarter-Monthly	13.162	11.720	4.196
Monthly	8.003	11.079	3.359

Table 14.8.3 shows the P-values for PPAR models for the three critical period statistics and for the six example series, while Table 14.8.4 contains the chi-squared values for the three critical period statistics with the series grouped according to their seasonal length. For the case of the monthly series, there are two P-values which suggest that the critical period statistics are preserved at the 1% level, but not at the 5% level. These relate to the extraction rate and deficiency percentage for the Saugeen Series. All other cases indicate preservation at the 5% level. The overall chi-squared test indicates the length of the critical period and deficiency percentages are preserved at the 5% level, while the extraction is preserved at the 1% level.

For the quarter-monthly series, evidence of preservation of the critical period statistics by PPAR models is not quite as strong as for the monthly series. Again, the extraction rate and deficiency for the Saugeen series are preserved at the 1% level, but not at the 5% level. The length of the critical period is not preserved at the 1% level for the Alcan system inflow series. All other statistics are preserved at the 5% level. Nevertheless, the overall chi-squared statistics indicate that the deficiency and the length of the critical period are preserved at the 5% level, while the extraction rate is preserved at the 1% level (and very close to being preserved at the 5% level).

Table 14.8.3. P-values for the PPAR models.

Riverflow Series		Statistics		
		Extraction	Length of CP	Deficiency
Quarter-Monthly	Alcan System	0.395	0.007	0.886
	Rio Grande	0.414	0.684	0.307
	Saugeen	0.011	0.767	0.035
Monthly	Alcan System	0.673	0.293	0.800
	Rio Grande	0.175	0.373	0.400
	Saugeen	0.012	0.799	0.034

Table 14.8.4. Chi-squared values for the PPAR models.

Seasonal Lengths	Statistics		
	Extraction	Length of CP	Deficiency
Quarter-Monthly	12.641	11.214	9.309
Monthly	13.124	4.876	9.042

It should be noted that in the majority of these simulation experiments (22 out of 36 combinations of models, series and critical period statistics), the coefficient of skewness of the empirical distribution of the critical period statistics is different from zero at the 5% level. In fact, for the length of critical period statistic, the skewness coefficient is always significantly different from zero at the 0.1% level. In view of the significant skewness encountered in this study, the types of statistical tests used by Hall et al. (1969) and Askew et al. (1971) are not appropriate. Their tests are based on the assumption of normality, and this assumption is not valid for skewed statistics.

A further point that should be stressed is that the split sample approach to testing the preservation of critical period statistics is more exact than the approach in which an entire series is used for both model fitting and the calculation of the statistics to be preserved. This latter approach was used in the earlier studies of Hall et al. (1969), and Askew et al. (1971), as well as in Section 10.6.4 for the Hurst statistics.

14.9 CONCLUSIONS

Because the basic mathematical design of the periodic models described in this chapter closely reflects the statistical characteristics of many kinds of seasonal time series, especially those arising in the environmental sciences, periodic models are ideally suited for use in practical applications. Of particular import is the family of PAR models for which comprehensive model construction techniques have been developed. If the number of model parameters has to be reduced, one can employ the economical PPAR class of models. Although more research is required to devise estimation algorithms for PARMA models that are computationally efficient, good progress has been made in the practical development of this promising class of models.

As just noted, periodic models are well designed for use with natural time series. When dealing with seasonal socio-economic time series in which the mean level and possibly other statistics may change within each season over the years, one may wish to experiment with the following modelling approach. Firstly, one can model the seasonal series, such as monthly water demand, using a SARIMA model (Chapter 12). Secondly, one can model the residuals of the fitted SARIMA model using a PAR or other type of periodic model. In this way, one may be able to model a seasonally varying correlation structure which is not captured by the SARIMA model.

The simulation experiments of Section 14.8.2, demonstrate that properly fitted PAR and PPAR models can preserve statistically important historical statistics. In the next chapter, it is shown using forecasting experiments that these periodic models forecast seasonal riverflow series better than both deseasonalized (Chapter 13) and SARIMA (Chapter 12) models.

PROBLEMS

14.1 Complete the following:

(a) Assuming that there are four seasons per year, and the order of the AR model in each season is two, write down the complete set of equations.

(b) Develop the theoretical periodic autocovariance function for the PAR model in part (a).

(c) Determine the periodic Yule-Walker equations for this model.

14.2 The stationarity requirement for the PAR model in [14.2.4] having one AR parameter in each season is given in [14.2.6]. By referring to appropriate references, determine the stationarity condition for a general PAR model that is not restricted to being Markov.

14.3 Complete the following:

(a) Using the notation in [14.2.15], write down the complete set of equations for a PARMA model having four seasons where $p_m = q_m = 1$ in the first two seasons, and $p_m = 2$ and $q_m = 1$ for the second two seasons.

(b) Derive the theoretical periodic autocovariance function for the PARMA model in part (a).

(c) Ascertain the periodic Yule-Walker equations for the model.

14.4 The stationarity and invertibility conditions for a PARMA model having one AR and one MA parameter in each season are given in [14.2.6] and [14.2.17], respectively. Present and explain the conditions for stationarity and invertibility for the general PARMA model in [14.2.15] for which the number of AR and MA parameters for a specific season are not restricted in number.

14.5 Suppose that a PARMA model for season m is given as

$$(1 - 0.11B + 0.30B^2)z_{r,m} = (1 - 0.4B)a_{r,m}$$

where the mean of $z_{r,m}$ is zero.

(a) Obtain the random shock coefficients for at least eight terms and then write this model in random shock form as in [14.2.24].

(b) Also write the model in inverted form as in [14.2.26].

14.6 Carry out the instructions in problem 14.5 for the PARMA model in season m which is given as

$$(1 - 0.10B + 0.24B^2)z_{r,m} = (1 - 0.4B)a_{r,m}$$

where the mean of $z_{r,m}$ is assumed to be zero.

14.7 Select an average monthly riverflow series and then fit a PAR model to this series adhering to the following steps in model construction:

(a) Examine appropriate exploratory data analysis graphs as well as the sample periodic ACF and PACF plots to design the most appropriate set of PAR models.

(b) Estimate the model parameters for each model selected in part (a) and then use the MAICE procedure to find the best one. For the most appropriate model compare the estimates for the model parameters employing both multiple linear regression and the periodic Yule-Walker equations. Comment upon the results.

(c) Carry out diagnostic checks to ensure that the best PAR model from (b) satisfies the whiteness, normality, and constant variance assumptions. If there are any problems make suitable modifications based upon the diagnostic results and repeat steps (b) and (c). Whatever the case, be sure to employ the periodic RACF test for whiteness given in [14.3.10].

14.8 Develop PAR models for describing average monthly riverflows from three distinctly different geographical locations in the world. Using identification results such as the sample periodic ACF and the sample periodic PACF graphs as well as the structures of the calibrated PAR models, make comparisons among the fitted models. Wherever appropriate, provide physical explanations as to why certain modelling results vary across the regions.

14.9 Develop the most appropriate PAR and PPAR models to describe an average monthly hydrological time series. Explain why any groupings of months within the PPAR model make sense or else do not seem reasonable from both statistical and physical viewpoints.

14.10 Follow the instructions in problem 14.9 for an average weekly hydrological time series.

14.11 Select a hydrological data set for which you have both average weekly and average monthly observations. Carry out the studies put forward in the previous two questions for the monthly and weekly time series. Subsequently, compare the monthly and weekly modelling results for the PAR and PPAR models. Did you find, for instance, that there were more groupings of seasons for the fitted weekly PPAR model than with the monthly version?

14.12 In Section 14.7, model building procedures are discussed for PARMA models. Summarize and compare according to both advantages and disadvantages the PARMA estimation techniques given by Vecchia (1985a,b) and Jimenez et al. (1989).

14.13 After fitting a PAR model to an average monthly riverflow time series, execute a proper simulation study to ascertain whether or not the historical critical period statistics given in [14.8.1] are preserved.

14.14 Fit a PAR model to a seasonal hydrological time series of your choice. Then carry out simulation experiments to determine if the sample periodic ACF in [14.3.4] at lag one for each season is preserved statistically by the calibrated model.

REFERENCES

CRITICAL PERIOD STATISTICS

Askew, A. J., Yeh, W. W. G. and Hall, W. A. (1971). A comparative study of critical drought simulation. *Water Resources Research*, 7:52-62.

Hall, W. A., Askew, A. J. and Yeh, W. W. G. (1969). Use of the critical period in reservoir analysis. *Water Resources Research*, 5(6):1205-1215.

McMahon, T. A. and Mein, R. G. (1978). *Reservoir Capacity and Yield*. Elsevier, New York.

Stedinger, J. R. and Taylor, M. R. (1982a). Synthetic streamflow generation 1.: Model verification and validation. *Water Resources Research*, 18(4):909-918.

Stedinger, J. R. and Taylor, M. R. (1982b). Synthetic streamflow generation 2.: Effect of parameter uncertainty. *Water Resources Research*, 18(4):919-924.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1987). Simulation of monthly hydrological time series. In MacNeill, I. B. and Umphrey, G. J., Editors, *Advances in Statistical Sciences*, Festschrift in Honor of Professor V. M. Joshi's 70th Birthday, volume IV, Stochastic Hydrology, pages 57-71, D. Reidel Publishing Co., Dordrecht, the Netherlands.

DATA SETS

Environment Canada (1977). Historical streamflow summary, Ontario. Technical report, Inland Waters Directorate, Water Resources Branch, Ottawa, Canada.

Thompstone, R. M., Poiré, A. and Valleé, A. (1980). A hydrometeorological information system for water resources management. *INFOR*, 18(3):258-274.

PERIODIC MODELLING

Bartolini, P., Salas, J. D. and Obeysekera, J. T. B. (1988). Multivariate periodic ARMA(1,1) processes. *Water Resources Research*, 24(8):1237-1246.

Cipra, T. (1985a). Periodic moving average processes. *Aplikace Matematily*, 30:218-229.

Cipra, T. (1985b). Statistical analysis of multiple moving average processes under periodicity. *Kybernetica*, 21:335-345.

Cipra, T. and Tlusty, P. (1987). Estimation of multiple autoregressive-moving average models using periodicity. *Journal of Time Series Analysis*, 8:293-301.

Claps, P., Rossi, F. and Vitale, C. (1993). Conceptual-stochastic modeling of seasonal runoff using autoregressive moving average models and different scales of aggregation. *Water Resources Research*, 29(8):2545-2559.

Croley II, T. E. and Rao, K. N. R. (1977). A manual for hydrologic time series deseasonalization and serial independence reduction. Report No. 199, Iowa Institute of Hydraulic Research, The University of Iowa, Iowa City, Iowa.

Dunsmuir, W. (1981). Estimation of periodically varying means and standard deviations in time series. *Journal of Time Series Analysis*, 2(3):129-153.

Fernandez, B. and Salas, J. D. (1986). Periodic gamma autoregressive processes for operational hydrology. *Water Resources Research*, 22(10):1385-1396.

Gladyshev, E. G. (1961). Periodically correlated random sequences. *Soviet Mathematics*, 2:385-388.

Gladyshev, E. G. (1963). Periodically and almost-periodically correlated random prrocesses with a continuous time parameter. *Theory of Probability and its Applications*, 8:173-177.

Haltiner, J. P. and Salas, J. D. (1988). Development and testing of a multivariate, seasonal ARMA(1,1) model. *Journal of Hydrology*, 104:247-272.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association*, 74(367):652-660.

Hurd, H. L. and Gerr, N. L. (1991). Graphical methods for determining the presence of periodic correlation. *Journal of Time Series Analysis*, 12:337-350.

Jimenez, C., McLeod, A. I. and Hipel, K. W. (1989). Kalman filter estimation for periodic autoregressive-moving average models. *Stochastic Hydrology and Hydraulics*, 3:227-240.

Jones, R. H. and Brelsford, W. M. (1967). Time series with periodic structure. *Biometrika*, 54:403-408.

Lewis, P. A. W. (1985). Some simple models for continuous variate time series. *Water Resources Bulletin*, 21(4):635-644.

Li, W. K. and Hui, Y. V. (1988). An algorithm for the exact likelihood of periodic autoregressive moving average models. *Communication in Statistics, Simulation and Computation*, 17(4):1483-1494.

McClave, J. T. (1978). Estimating the order of autoregressive models: The Max χ^2 method. *Journal of the American Statistical Association*, 73(363):122-128.

McLeod, A. I. and Hipel, K. W. (1978). Developments in monthly autoregressive modelling. Technical Report 45-XM-011178, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario.

Morgan, J. A. and Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14(2):317-325.

Moss, M. E. and Bryson, M. C. (1974). Autocorrelation structure of monthly streamflows. *Water Resources Research*, 10:737-744.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika*, 61(3):423-426.

Newton, H. J. (1982). Using periodic autoregressions for multiple spectral estimation. *Technometrics*, 24:109-116.

Noakes, D. J., McLeod, A. I. and Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, 1:179-190.

Obeysekera, J. T. B. and Salas, J. D. (1986). Modeling of aggregated hydrologic time series. *Journal of Hydrology*, 86:197-219.

Pagano, M. (1978). On periodic and multiple autoregressions. *The Annals of Statistics*, 6(6):1310-1317.

Rose, D. E. (1977). Forecasting aggregates of independent ARIMA processes. *Journal of Econometrics*, 5:323-345.

Sakai, H. (1982). Circular lattice filtering using Pagano's method. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-30(2):279-287.

Salas, J. D. and Abdelmohsen, M. W. (1993). Initialization for generating single-site and multisite low-order periodic autoregressive and moving average processes. *Water Resources Research*, 29(6):1771-1776.

Salas, J. D., Boes, D. C. and Smith, R. A. (1982). Estimation of ARMA models with seasonal parameters. *Water Resources Research*, 18(4):1006-1010.

Salas, J. D., Delleur, J. W., Yevjevich, V. and Lane, W. L. (1980). Applied modelling of hydrologic series. *Water Resources Publications*, Littleton, Colorado.

Salas, J. B. and Obeysekera, J. T. B. (1992). Conceptual basis of seasonal streamflow time series models. *Journal of Hydraulic Engineering*, ASCE, 118(8):1186-1194.

Salas, J. D., Tabios III, G. Q. and Bartolini, P. (1985). Approaches to multivariate modeling of water resources time series. *Water Resources Bulletin*, 21(4).

Tao, P. C. and Delleur, J. W. (1976). Seasonal and nonseasonal ARMA models. *Journal of the Hydraulics Division, ASCE*, 102(HY10):1541-1559.

Thomas, H. A. and Fiering, M. B. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Maass, A., Hufshmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A. and Fair, M. G., Editors, *Design of Water Resources Systems*, pages 459-493. Harvard University Press.

Thompstone, R. M. (1983). *Topics in Hydrological Time Series Modelling*. PhD thesis, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985a). Grouping of periodic autoregressive models. In Anderson, O. D., Ord, J. K. and Robinson, E. A., Editors, *Time Series Analysis: Theory and Practice 6*, pages 35-49. North-Holland, Amsterdam.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985b). Forecasting quarter-monthly riverflow. *Water Resources Bulletin*, 21(5):731-741.

Tiao, G. C. and Gruppe, M. R. (1980). Hidden periodic autoregressive-moving average models. *Biometrika*, 67(2):365-373.

Troutman, B. M. (1979). Some results in periodic autoregression. *Biometrika*, 66(2):219-228.

Ula, T. A. (1990). Periodic covariance stationarity of multivariate periodic autoregressive moving average processes. *Water Resources Research*, 26(5):855-861.

Vecchia, A. V. (1985a). Periodic autoregressive-moving average (PARMA) modelling with applications to water resources. *Water Resources Bulletin*, 21(5):721-730.

Vecchia, A. V. (1985b). Maximum likelihood estimation for periodic autoregressive-moving average models. *Technometric*, 27:375-384.

Vecchia, A. V. and Ballerini, R. (1991). Testing for periodic autocorrelations in seasonal time series data. *Biometrika*, 78(1):53-63.

Vecchia, A. V., Obeysekera, J. T., Salas, J. D., and Boes, D. C. (1983). Aggregation and estimation for low-order periodic ARMA models. *Water Resources Research*, 19(5):1297-1306.

PORTMANTEAU TEST STATISTIC

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised edition.

Box, G. E. P. and Pierce, D. A. (1970). Distribution of the residual autocorrelations in autoregressive integrated moving average models. *Journal of the American Statistical Association*, 65:1509-1526.

Davies, N., Triggs, C. M. and Newbold, P. (1977). Significance of the Box-Pierce Portmanteau statistics in finite samples. *Biometrika*, 64:517-522.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:297-303.

McLeod, A. I. (1993). Diagnostic checking periodic autoregression models with application. *Journal of Time Series Analysis*.

STATISTICS

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29:610-611.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburg, England.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburg, England.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31:9-12.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York, second edition.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. The Iowa State University Press, Ames, Iowa.

Wickmann, B. A. and Hill, I. D. (1982). An efficient and portable pseudo-random number generator. *Applied Statistics*, 31:188-190.

CHAPTER 15

FORECASTING

WITH

SEASONAL MODELS

15.1 INTRODUCTION

Three families of models are presented in Part VI of the book for fitting to seasonal time series. In particular, SARIMA, deseasonalized and periodic models are described in Chapters 12 to 14, respectively. The objective of this chapter is to employ *forecasting experiments* for comparing the capabilities of these seasonal models to forecast accurately seasonal hydrological time series.

Forecasting can be utilized for *model discrimination* purposes. After fitting different types of models to one or more time series by following proper model construction procedures, the model or models which forecasts the best according to certain criteria can be selected for use in further practical applications. Because carrying out forecasting studies is a very time consuming undertaking, forecasting experiments cannot be used for discriminating among models in most applications. Nonetheless, if one finds, for example, in an extensive forecasting experiment, that a certain type of PAR model forecasts significantly better than its competitors when used with average monthly riverflow series, this would give one confidence in using PAR models in other applications involving average monthly riverflow series.

After explaining *how to calculate forecasts* for seasonal models in Section 15.2, two main forecasting studies are described in the next two major sections. In the *first set of forecasting experiments*, mean monthly flows from thirty rivers in North and South America are used to test the short-term forecasting ability of SARIMA, deseasonalized and PAR models. After splitting each series into two sections, the seasonal models are calibrated for the first portion of the data. The fitted models are then used to generate one-step ahead forecasts for the second portion of each time series. The forecasting performance of the models is compared using various measures of accuracy. The results suggest that PAR models identified using the sample periodic ACF and PACF provide the most accurate forecasts. The results of this study are also presented by Noakes et al. (1985) as well as Noakes (1984, Ch. V).

In the *second forecasting study*, the three quarter-monthly and three monthly riverflow series used in Sections 14.6 and 14.8.2 of the previous chapter, are used for comparing the forecasting accuracy of seasonal models. Besides the SARIMA (Chapter 12), deseasonalized (Chapter 13) and PAR models (Sections 14.2.2, 14.3, 14.4, 14.6 and 14.8), the PPAR models (Sections 14.5, 14.6 and 14.8) are also employed in this forecasting experiment. This second forecasting study was originally presented by Thompstone (1983, Ch. 4).

In both sets of forecasting experiments, *one step ahead forecasts* are used for comparing the forecasting abilities of the model. There are two reasons for doing this. Firstly, from a theoretical viewpoint one can show that for the families of seasonal models presented in

Chapters 12 to 14, the one-step ahead forecasts are independent of one another. This property allows one to use *statistical tests* based upon the independence assumption to ascertain whether or not one model forecasts significantly better than another. Secondly, in many practical applications the one step ahead forecasts are of most importance to decision makers. For example, when deciding upon the operating rules of a reservoir for generation of hydro-electric power, an accurate forecast for the inflows of the next month is crucial. After the real value of next month's flows is known, one can use this information in the seasonal forecasting model to produce the one step ahead forecast for the subsequent month and so on.

Because different kinds of time series models are not defined and calibrated in exactly the same way, it is not surprising that their forecasts for a given time series are not identical. In fact, a given type of model approaches forecasting from a unique perspective based upon its own particular strengths and weaknesses. To attempt to exploit the forecasting capabilities of each kind of model fitted to a time series, forecasts generated by individual models can be combined in an optional manner. Procedures for combining forecasts across models are presented in Section 15.5.2. Additionally, experimental results on combining forecasts for SARIMA and PAR models fitted to average monthly riverflows are given in Section 15.5.3, while findings on combining hydrological forecasts from transfer function-noise (TFN), PAR and conceptual models are described in Section 18.4.2.

Before the conclusions, a brief discussion is given in Section 15.6 on aggregating forecasts for the purpose of producing a forecast for a longer time interval. For instance, one can employ a monthly model to produce 12 monthly forecasts and then sum these 12 values to obtain the aggregated annual forecast.

For a summary of where material on forecasting is presented in the book, the reader can refer to Table 1.6.3. In particular, the table points out that forecasting with nonseasonal ARMA and TFN models is described in Chapters 8 and 18, respectively. Finally, for references on forecasting listed outside of this chapter, the reader may wish to refer to appropriate references given at the end of Chapter 1 as well as Chapters 8 and 18.

15.2 CALCULATING FORECASTS FOR SEASONAL MODELS

15.2.1 Introduction

Suppose that one fits an appropriate seasonal model to a seasonal time series and then wishes to forecast l steps ahead where $l \geq 1$. When using z_t to represent the value of the time series, as is done in Chapter 12 with SARIMA models, one can employ the calibrated seasonal model to forecast z_{t+l} given the observations up to and including time t. As explained in Section 8.2 for nonseasonal ARMA models, the *minimum mean square error (MMSE) forecast* $\hat{z}_t(l)$ for z_{t+l} can be obtained by minimizing $E[z_{t+l} - \hat{z}_t(l)]^2$. This minimization is equivalent to taking the conditional expectation of z_{t+l} at time t.

For the deseasonalized and periodic models of Chapters 13 and 14, respectively, it is convenient to let $z_{r,m}$ stand for the observation in year r and season m. Then $\hat{z}_{r,m}(l)$ represents the MMSE forecast for lead time $l \geq 1$ starting at $z_{r,m}$.

The general approach for calculating MMSE forecasts for the seasonal models of Part VI is very similar to that used for nonseasonal ARMA models in Section 8.2. The specific method for calculating MMSE forecasts for each of the seasonal models is described below.

15.2.2 Forecasting with SARIMA Models

The SARIMA model is defined in [12.2.7]. The most convenient format to employ when calculating MMSE forecasts is the generalized form of the SARIMA model given in [12.2.12]. More specifically, to calculate the conditional expectation of z_{t+l} at time t and, hence, the MMSE forecast $\hat{z}_t(l)$, one takes conditional expectations of [12.2.12] to obtain

$$\begin{aligned}[z_{t+l}] &= \phi'_1[z_{t+l-1}] + \phi'_2[z_{t+l-2}] + \cdots \\ &\quad + \phi'_{p+sP+d+sD}[z_{t+l-p-sP-d-sD}]+[a_{t+l}] - \theta'_1[a_{t+l-1}] \\ &\quad - \theta'_2[a_{t+l-2}] - \cdots - \theta'_{q+sQ}[a_{t+l-q-sQ}]\end{aligned} \qquad [15.2.1]$$

where

$l = 1,2, \ldots,$ is the lead time for the forecast,

$[z_{t+l}]$ denotes the conditional expectation

$\underset{t}{E}[z_{t+l}|z_t,z_{t-1}, \cdots];$

ϕ'_i is the generalized AR parameter defined by

$\phi'(B) = \phi(B)\Phi(B^s)\nabla^d\nabla_s^D$; and

θ'_i is the generalized MA parameter defined by

$\theta'(B) = \theta(B)\Theta(B^s).$

The nonseasonal version of [15.2.1] is given in [8.2.22]. As explained in Section 8.2.4 for forecasting with a nonseasonal ARMA model, one can allow for a nonzero deterministic trend component by introducing the parameter θ_0 on the right hand side of [8.2.21] to obtain [8.2.23]. By taking conditional expectations of [8.2.23], one obtains [8.2.24] for calculating MMSE forecasts for a nonseasonal ARMA model containing the level parameter θ_0. In a similar fashion for a SARIMA model, one can introduce the parameter θ_0 on the right hand side of [12.2.12] and then take conditional expectations to obtain a formula for calculating MMSE forecasts. The resulting formula would be the same as [15.2.1] expect for the parameter θ_0 which would be added to the right hand side.

As is also done in Section 8.2.4 for nonseasonal ARMA models, the conditional expectations in [15.2.1] can be determined using the following four rules:

(1) $$\underset{t}{E}[z_{t-j}] = z_{t-j}, \; j = 0,1,2, \ldots, \qquad [15.2.2]$$

(2) $$\underset{t}{E}[z_{t+j}] = \hat{z}_t(j), \; j = 0,1,2, \ldots, \qquad [15.2.3]$$

is the MMSE forecast for lead time j,

(3) $$\underset{t}{E}[a_{t-j}] = a_{t-j},\ j = 0,1,2,\ldots, \qquad [15.2.4]$$

and

(4) $$\underset{t}{E}[a_{t+j}] = 0,\ j = 1,2,\ldots. \qquad [15.2.5]$$

If the series contains a level represented by θ_0, this can be added to the forecasts obtained using the above rules.

The MMSE forecasts have a number of interesting properties which can be illustrated using the random shock form of the model in [12.2.15]. The forecast at time t for lead time l is

$$\hat{z}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots \qquad [15.2.6]$$

Subtracting this from z_{t+l}, the forecast error is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \qquad [15.2.7]$$

Since $E[e_t(l)] = 0$, the variance of the forecast error is

$$V(l) = [Var\ e_t(l)] = [1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2]\sigma_a^2 \qquad [15.2.8]$$

This variance can be utilized to estimate confidence intervals for forecasts at various lead times.

The one step ahead forecast error is

$$e_t(1) = z_{t+1} - \hat{z}_t(1) = a_{t+1} \qquad [15.2.9]$$

Although one step ahead forecast errors are statistically independent, forecast errors for lead times greater than one are correlated. For forecasts made from origin t, the correlation coefficient between forecast errors at lead times l and $l + j$ is given as (Box and Jenkins, 1976)

$$\rho[e_t(l), e_t(l+j)] = \frac{\sum_{i=0}^{l-1} \psi_i \psi_{j+i}}{\left\{\sum_{h=0}^{l-1} \psi_h^2 \sum_{g=0}^{l+j-1} \psi_g^2\right\}^{1/2}} \qquad [15.2.10]$$

Inverse Box-Cox Transformation

Often the given series, z_t, is first transformed using the Box-Cox transformation in [12.2.1] to obtain the $z_t^{(\lambda)}$ series. The SARIMA model is then fitted to the $z_t^{(\lambda)}$ series as in [12.2.7]. The above calculations for obtaining MMSE forecasts are then carried out for the $z_t^{(\lambda)}$ series rather than z_t.

A naive approach for obtaining forecasts in the untransformed domain is to take the inverse Box-Cox transformation of the MMSE forecasts calculated in the transformed domain. However, in order to produce MMSE forecasts in the untransformed domain, a modified type of inverse Box-Cox transformation must be employed. More specifically, the exact MMSE forecast in the untransformed domain is determined from the fact that its transformed value follows a normal distribution with expected value $\hat{z}_t^{(\lambda)}(l)$ and variance $V(l)$. The expected value of the inverse Box-Cox transformed value is the desired MMSE forecast and it is determined numerically by Hermite polynomial integration (Granger and Newbold, 1976). In practice, it is found that the MMSE forecasts are slightly smaller than the naive forecasts. Moreover, studies with real data have shown that these MMSE forecasts do perform better than the naive forecasts. When data are transformed using a natural logarithmic transformation, as is often the case for seasonal hydrological time series, the MMSE forecast for the untransformed data is

$$\hat{z}_t(l) = exp[\hat{z}_t^{(\lambda)}(l) + \frac{V(l)}{2}] - c \ , \quad l = 1,2,\ldots, \qquad [15.2.11]$$

where $\hat{z}_t(l)$ is the MMSE forecast in the untransformed domain, $\hat{z}_t^{(\lambda)}(l)$ is the MMSE forecast produced by the model for the transformed logarithmic data, $V(l)$ is the variance of the forecast error given in [15.2.8], and c is the constant in the Box-Cox transformation required to make all entries be greater than zero.

For graphs of forecasts obtained using SARIMA models fitted to seasonal time series, the reader can refer to Section 12.5. In particular, Figures 12.5.1 and 12.5.2 display MMSE forecasts for monthly water demands and concentrations of atmospheric CO_2, respectively.

15.2.3 Forecasting with Deseasonalized Models

The main steps involved in forecasting with deseasonalized models are displayed in Figure 13.5.1. Firstly, one must calculate the MMSE forecasts for the ARMA model fitted to the deseasonalized series. This procedure is identical to that presented for the nonseasonal ARMA(p,q) model in Section 8.2. Let the deseasonalized series that is determined using either [13.2.2] or [13.2.3] be represented as $w_{r,m}$, where r and m stand for the year and season, respectively. By taking conditional expectations of the ARMA(p,q) model in [13.2.12], the MMSE forecasts for the deseasonalized series are calculated using

$$\begin{aligned}[w_{r,m+l}] &= \phi_1[w_{r,m+l-1}]+\phi_2[w_{r,m+l-2}] + \cdots + \phi_p[w_{r,m+l-p}] + [a_{r,m+l}] \\ &\quad - \theta_1[a_{r,m+l-1}] - \theta_2[a_{r,m+l-2}] - \cdots - \theta_q[a_{r,m+l-q}] \qquad [15.2.12]\end{aligned}$$

where

$l = 1,2,\ldots,$ is the lead time for the forecast, and

$[w_{r,m+l}]$ denotes the conditional expectation

$$\underset{t}{E}[w_{r,m+l}|w_{r,m}, w_{r,m-1},\ldots,] \ .$$

Equation [15.2.12] can be used to calculate MMSE forecasts for the deseasonalized series by following the four rules given below for $l = 1,2,\ldots,$

(1) $$\underset{t}{E}[w_{r,m-j}] = w_{r,m-j}, \ j = 0,1,2,\ldots, \qquad [15.2.13]$$

(2) $$\underset{t}{E}[w_{r,m+j}] = \hat{w}_{r,m}(j), \ j = 1,2,\ldots, \qquad [15.2.14]$$

is the MMSE forecast for $w_{r,m+j}$,

(3)
$$\underset{t}{E}[a_{r,m-j}] = a_{r,m-j},\ j = 0,1,2,\ldots, \qquad [15.2.15]$$

and

(4)
$$\underset{t}{E}[a_{r,m+j}] = 0,\ j = 1,2,\ldots. \qquad [15.2.16]$$

When using the above rules, one should keep in mind that the time of occurrence of the deseasonalized series or the innovations can be written using a variety of equivalent subscripts. For instance, when there are s seasons per year $w_{r,m}$, $w_{r-1,m+s}$ and $w_{r+1,m-s}$ all stand for the same value.

Following the procedure described in Section 3.4.3, the random shock coefficient, ψ_i, $i = 1,2,\ldots,$ can be found for the ARMA(p,q) model describing the $w_{r,m}$ series. The variance of the forecast error for the deseasonalized series can then be determined as

$$V(l) = [1 + \psi_i^2 + \psi_2^2 + \cdots + \psi_{l-1}^2] \qquad [15.2.17]$$

Finally, the one step ahead forecast error is

$$e_t(1) = w_{r,m+1} - \hat{w}_{r,m}(1) = a_{r,m+1} \qquad [15.2.18]$$

As indicated in Figure 13.5.1, the next step is to route the MMSE forecasts through the inverse deseasonalization filter to obtain $z_{r,m}^{(\lambda)}$. The inverse deseasonalization for the two techniques given in [13.2.2] and [13.2.3] are

$$\hat{z}_{r,m}^{(\lambda)}(l) = \hat{w}_{r,m}(l) + \bar{\mu}_m \qquad [15.2.19]$$

and

$$\hat{z}_{r,m}^{(\lambda)}(l) = \hat{w}_{r,m}(l)\bar{\sigma}_m + \bar{\mu}_m, \qquad [15.2.20]$$

respectively. To obtain forecasts in the untransformed domain one must take the inverse Box-Cox transformation of $\hat{z}_{r,m}^{(\lambda)}(l)$. However, as noted in the previous subsection, if one wishes to have MMSE forecasts in the untransformed domain, one must make an appropriate adjustment before taking the inverse Box-Cox transformation. For the case of a logarithmic transformation, the MMSE forecast given in the same units as the original series is determined using

$$\hat{z}_{r,m}(l) = \exp[\hat{z}_{r,m}^{(\lambda)}(l) + \frac{1}{2}V(l)] - c \qquad [15.2.21]$$

where $V(l)$ is the variance of the forecast error from [15.2.17].

15.2.4 Forecasting with Periodic Models

In Section 8.2, it is explained how to calculate MMSE forecasts for a nonseasonal ARMA model. A similar procedure is followed when forecasting with PAR, PPAR or PARMA models. For example, when calculating MMSE forecasts for a PAR model, one simply writes down the difference equation for season m in [14.2.1] and then determines the conditional expectations of the observations and innovations to arrive at the MMSE forecasts. Likewise, for a PARMA model, one uses the difference equation in [14.2.15] for the ARMA model in season m and then calculates the conditional expectations.

The approach for calculating MMSE forecasts for PARMA models is explained first. Assuming that the observations and innovations are known up to the rth year and mth season, one takes the conditional expectation of [14.2.15] to obtain

$$[z_{r,m+l}^{(\lambda)}] = \phi_1^{(m)}[z_{r,m+l-1}^{(\lambda)}] + \phi_2^{(m)}[z_{r,m+l-2}^{(\lambda)}] + \cdots + \phi_{p_m}^{(m)}[z_{r,m+l-p_m}^{(\lambda)}] + [a_{r,m}]$$

$$- \theta_1^{(m)}[a_{r,m+l-1}] - \theta_2^{(m)}[a_{r,m+l-1}] - \cdots - \theta_{q_m}^{(m)}[a_{a_{r,m+l-q_m}}] \qquad [15.2.22]$$

where

$l = 1,2, \ldots,$ is the lead time for the forecast, and

$[z_{r,m+l}^{(\lambda)}]$ denotes the conditional expectation

$$\underset{t}{E}[z_{r,m+l}^{(\lambda)} | z_{r,m}^{(\lambda)}, z_{r,m-1}^{(\lambda)}, \ldots]$$

By following the four rules listed below, equation [15.2.22] can be employed for calculating the MMSE forecasts for $z_{r,m}^{(\lambda)}$ for lead times $l = 1,2, \ldots,$.

(1) $$\underset{t}{E}[z_{r,m-j}^{(\lambda)}] = z_{r,m-j}^{(\lambda)}, \; j = 0,1,2, \ldots, \qquad [15.2.23]$$

(2) $$\underset{t}{E}[z_{r,m+j}^{(\lambda)}] = \hat{z}_{r,m}^{(\lambda)}(j), \; j = 1,2, \ldots, \qquad [15.2.24]$$

is the MMSE forecast for $z_{r,m+j}^{(\lambda)}$,

(3) $$\underset{t}{E}[a_{r,m-j}] = a_{r,m-j}, \; j = 0,1,2, \ldots, \text{ and} \qquad [15.2.25]$$

(4) $$\underset{t}{E}[a_{r,m+j}] = 0, \; j = 1,2, \ldots, \; . \qquad [15.2.26]$$

After calculating the forecasts for $l = 1,2, \ldots,$ the appropriate monthly mean μ_m must be added to each forecast when $\mu_m \neq 0$.

The procedure in Section 3.4.3 can be utilized to find the random shock coefficients $\psi_i^{(m)}$, $i = 1,2, \ldots,$ for the ARMA model in season m. To calculate the variance of the forecast error for $\hat{z}_{r,m}^{(\lambda)}(l)$ one uses

$$V^{(m)}(l) = [1 + \psi_1^{(m)^2} + \psi_2^{(m)^2} + \cdots + \psi_{l-1}^{(m)^2}]\sigma_m^2 \qquad [15.2.27]$$

The one step ahead forecast error can be shown to be

$$e_t^{(m)}(1) = z_{r,m+1}^{(\lambda)} - \hat{z}_{r,m}^{(\lambda)}(1) = a_{r,m+1} \qquad [15.2.28]$$

To obtain forecasts in the same units as the original series, one must take the inverse Box-Cox transformation of $\hat{z}_{r,m}^{(\lambda)}(l)$ for $l = 1,2, \ldots,$. When the data are transformed using natural logarithms (i.e., $\lambda = 0$), equation [15.2.21] can be utilized to calculate the MMSE forecasts in the untransformed domain where $V(l)$ is replaced by $V^{(m)}(l)$ from [15.2.27].

When determining forecasts for PAR or PPAR models, one can follow the approach explained for PARMA models. Consider, for example, the case of the PAR model. By taking conditional expectations of [14.2.1] or [14.2.3], the MMSE forecasts calculated after year r and season m are determined using

$$[z_{r,m+l}^{(\lambda)}] = \phi_1^{(m)}[z_{r,m+l-1}^{(\lambda)}] + \phi_2^{(m)}[z_{r,m+l-2}^{(\lambda)}] + \cdots + \phi_{p_m}^{(m)}[z_{r,m+l-p_m}^{(\lambda)}] + [a_{r,m}] \qquad [15.2.29]$$

The four rules presented in [15.2.23] to [15.2.26] can then be used to calculate the MMSE forecasts for the transformed series. Additionally, when $\mu_m \neq 0$, one must add the appropriate mean level to each of the calculated forecasts. Finally, the modified version of the inverse Box-Cox transformation (see [15.2.21] for the case of $\lambda = 0$) must be taken to produce MMSE forecasts in the untransformed domain.

15.3 FORECASTING MONTHLY RIVERFLOW TIME SERIES

15.3.1 Introduction

To examine the efficacy of PAR models of Chapter 14, a comprehensive forecasting study is carried out by comparing their performance with that of several models used to model seasonal data. Using thirty monthly riverflow time series, the PAR models are compared to the SARIMA models of Chapter 12 as well as the deseasonalized models presented in Chapter 13. Methods of model order selection for the PAR models are also compared. The experiments described in this section, as well as by Noakes et al. (1985), are the most comprehensive yet reported in the hydrological literature. Other published comparisons have used only a few series and usually only two models [see, for example, Delleur et al. (1976)]. Also, the majority of the hydrological forecasting research to date has been concentrated on shorter time intervals in the order of a few hours or days [see, for example, the Proceedings of the Oxford Hydrological Forecasting Symposium, April 15-18 (International Association of Hydrological Sciences, 1980) and Thompstone et al. (1983)]. However, monthly riverflow forecasts are often used for operational planning of reservoir systems. Camacho (1990) considers both short term and long term forecasts in his riverflow forecasting study. Even modest improvements in the operation of large reservoir systems can result in multi-million dollar savings per year (see, for instance, Brocha (1978) as well as the comments on stochastic hydrology given in Section 1.1). Thus, the results of the forecasting study given in this section should be important to those concerned with the optimal medium and long-term operation of reservoir systems.

The performance of the forecasts from the different seasonal models are assessed using the *root mean square error (RMSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE),* and *median absolute percentage error (MEDIAN APE),* criteria. Although these criteria give an indication as to which models seem to perform better, no statement concerning statistically significant differences can be made from such a comparison. To address this question, the nonparametric Wilcoxon signed rank test (Wilcoxon, 1945) is used to determine if a particular model produces significantly better forecasts when compared to another model. One could also employ Pitman's (1939) correlation test and the likelihood ratio test to check if one model forecasts significantly better than another. These latter two tests are described in Section 8.3.2 and used in the forecasting experiments with nonseasonal models presented in Section 8.3.4. The nonparametric Wilcoxon test is outlined in this section with the seasonal forecasting experiments and described in detail in Appendix A23.2. Noakes et al. (1983) and Noakes

(1984) present the results of the forecasting study of this section when Pitman's correlation study and the likelihood ratio tests are used. Finally, the overall procedure for carrying out the forecasting experiments in this section, Section 15.4 as well as Sections 8.3 and 15.3, is summarized in Figure 8.3.1.

15.3.2 Data Sets

The data used in this study comprise thirty monthly unregulated riverflow time series ranging in length from thirty-seven to sixty-eight years. The rivers are from a number of different physiographic regions and vary in size from a river with a mean annual flow of one cubic meter per second (m^3/s) to a river having a mean annual flow of almost 900 m^3/s. The data for the Canadian rivers were obtained from Water Survey of Canada records, the American riverflow series are from the United States Geological Survey, and the Brazilian data were kindly provided from Electrobras (the national electrical company of Brazil). The rivers and their mean annual flows for the water year from October to September are displayed in Table 15.3.1.

15.3.3 Seasonal Models

The last three years or 36 observations are omitted from each of the data sets in Table 15.3.1. After taking natural logarithms of the time series, SARIMA, deseasonalized and PAR models are fitted to the thirty truncated logarithmic series.

The most appropriate SARIMA models to fit to the series are identified using the graphical procedures of Section 12.3.2. All of the SARIMA models identified for fitting to the monthly riverflow series in Table 15.3.1 are determined to be of the form $(p,0,q)\times(0,1,Q)_{12}$ with $\lambda = 0$ and with typical values of p, q and Q being 1, 0 and 1.

Two types of deseasonalized models are used in the forecasting study. For the first kind of model, equation [13.2.2] is used to deseasonalize the logarithmic data after estimating each monthly mean of the logarithmic data using [13.2.4]. The most appropriate ARMA model is then fitted to this deseasonalized series using the model construction techniques of Part III. This overall deseasonalized model is referred to as DSM.

For the second type of deseasonalized model, equation [13.2.3] is used to deseasonalize the logarithmic series before fitting an ARMA model to the resulting nonseasonal series. In [13.2.3], the seasonal means and standard deviations are estimated using [13.2.4] and [13.2.5], respectively. This overall deseasonalized model is called DES.

Six types of PAR models are considered in this study. In the first model, a separate AR(1) model is fitted to each month (called PAR/1) using multiple linear regression. This model was originally suggested by Thomas and Fiering (1962) and has been used extensively by hydrologists.

The second and third PAR models are fitted to the data using the algorithm of Morgan and Tatar (1972) described in Section 14.3.3. This algorithm calculates the residual sum of squares of all possible regressions for each season. The AIC and BIC can thus be calculated for all possible models. The PAR model which gives the minimum value of the AIC in [14.3.8] or BIC in [6.3.5] (with $p_m \leq 12$) is selected as the most appropriate. This type of procedure has been called subset autoregression by McClave (1975), and thus is referred to as SUBSET/AIC or SUBSET/BIC modelling.

Table 15.3.1. Average monthly riverflow time series used in the forecasting experiments.

	River	Location	Period	Observations	Mean Flow (m^3/s)
1	American	Fair Oaks, California	1906-1960	660	106
2	Boise	Twin Springs, Idaho	1912-1960	588	33
3	Clearwater	Kamish, Idaho	1911-1960	600	231
4	Columbia	Nicholson, British Columbia	1933-1969	444	109
5	Current	Van Buren, Missouri	1922-1960	468	54
6	W.B. Delaware	Hale Eddy, New York	1916-1960	540	30
7	English	Sioux Lookout, Ontario	1922-1977	660	123
8	Feather	Oroville, California	1902-1977	708	167
9	James	Buchanan, Virginia	1911-1960	600	69
10	Judith	Utica, Montana	1920-1960	492	1
11	Mad	Springfield, Ohio	1915-1960	552	14
12	Madison	West Yellowstone, Montana	1923-1960	456	13
13	McKenzie	McKenzie Bridge, Oregon	1911-1960	600	47
14	Middle Boulder	Nederland, Colorado	1912-1960	588	2
15	Missinaibi	Mattice, Ontario	1921-1976	672	103
16	Namakan	Lac La Croix, Ontario	1923-1977	648	108
17	Neches	Rockland, Texas	1914-1960	564	69
18	N. Magnetawan	Burke Falls, Ontario	1916-1977	732	6
19	Oostanaula	Resaca, Georgia	1893-1960	816	78
20	Pigeon	Middle Falls, Ontario	1924-1977	636	14
21	Rappahannock	Fredericksburg, Virginia	1908-1971	768	45
22	Richelieu	Fryers Rapids, Quebec	1932-1977	468	331
23	Rio Grande	Furnas, Minas Gerais, Brazil	1931-1978	576	896
24	Saint Johns	Fort Kent, New Brunswick	1927-1977	600	30
25	Saugeen	Walkerton, Ontario	1915-1976	744	68
26	S.F. Skykomish	Index, Washington	1923-1960	456	278
27	S. Saskatchewan	Saskatoon, Saskatchewan	1911-1963	624	272
28	Trinity	Lewiston, California	1912-1960	588	47
29	Turtle	Mine Centre, Ontario	1921-1977	672	37
30	Wolf	New London, Wisconsin	1914-1960	564	49

The next PAR models are estimated by using the appropriate Yule-Walker equations (see Section 14.3.3). In the first case p_m is selected on the basis of the minimum value of the AIC or BIC. Unlike the previous case, however, intermediate parameters are not allowed to be constrained to zero. Thus, all of the parameters from $\phi_1^{(m)}$ to $\phi_{p_m}^{(m)}$ are estimated in this model for a given season to produce the PAR/AIC and PAR/BIC models.

The last PAR models are identified by examining plots of the sample periodic PACF, presented in Section 14.3.2. In general, an AR(p_m) model is fitted to month m, where p_m is the last lag for which the PACF is significantly different from zero. The adequacy of the selected model is checked by testing for significant residual correlation or non-normality. Thus, the PAR/PACF is the natural extension to PAR models of the modelling philosophy recommended by Box and Jenkins (1976) and adhered to in this book. Once again, no intermediate parameters are constrained to zero.

15.3.4 Forecasting Study

After omitting the last 36 values of each of the 30 average monthly riverflow series in Table 15.3.1, the nine seasonal models are fitted to the 30 truncated series. From Section 15.3.3, these nine models are labelled as the SARIMA, DSM, DES, SUBSET/AIC, SUBSET/BIC, PAR/AIC, PAR/BIC, PAR/1, and PAR/PACF models. The nine models are then used to generate thirty-six one-step-ahead forecasts for the logarithmic flows. Figure 15.3.1 shows a time series plot of the last five years of the logarithmic flows along with the forecasts for the last three years using the PAR/PACF method for river number 14 in Table 15.3.1. As can be seen from a visual viewpoint, the PAR/PACF model forecasts quite well.

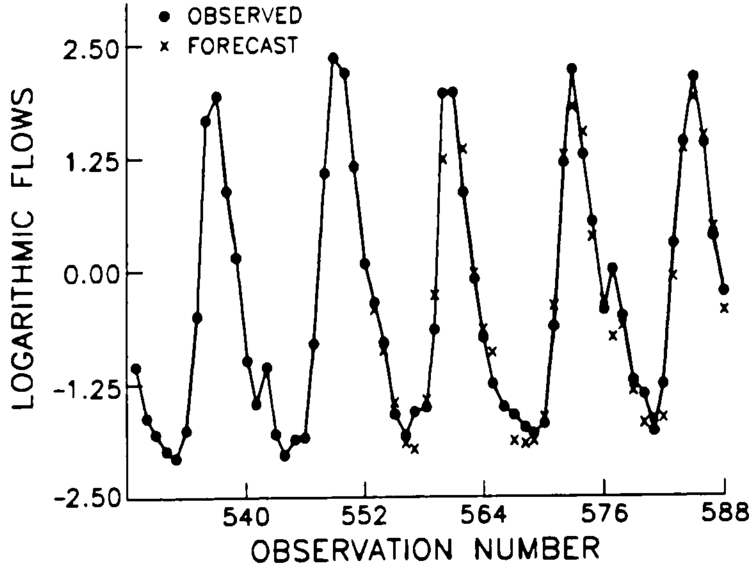


Figure 15.3.1. Logarithmic monthly flows and one step ahead PAR/PACF forecasts for the Middle Boulder Creek.

The monthly means of the logarithmic flows are also considered as forecasts and are referred to as MEANS. The logarithmic forecast errors associated with each of the ten forecasting models are then compared using the forecast performance measures RMSE, MAD, MAPE and MEDIAN APE, mentioned in Section 15.3.1.

RMSE results are given in Table 15.3.2 for each river. The results for each performance measure are summarized in Tables 15.3.3 to 15.3.6 where rank and rank-sum comparisons appear.

Table 15.3.2. RMSE × 1000 of the logarithmic forecast errors.

River	PAR/ PACF	PAR/1	PAR/ AIC	PAR/ BIC	SUBSET/ AIC	SUBSET/ BIC	DSM	DES	SARIMA	MEANS
1	857	896	813	864	796	796	801	907	690	1240
2	280	279	273	280	307	289	264	289	273	248
3	323	330	334	331	346	330	359	339	367	544
4	183	190	180	198	204	211	184	181	182	209
5	426	418	445	410	464	423	389	408	390	357
6	658	642	628	666	681	664	689	690	698	775
7	191	218	187	203	218	201	209	205	440	633
8	337	338	394	338	415	335	354	347	358	481
9	516	495	536	544	562	548	489	489	488	579
10	470	469	463	469	500	471	582	427	576	746
11	435	428	416	431	481	440	426	431	424	539
12	98	91	120	90	125	98	98	118	107	127
13	175	175	208	176	254	221	167	171	169	186
14	273	273	272	274	281	296	290	290	302	365
15	619	614	604	618	634	626	707	639	752	961
16	242	244	238	243	248	238	253	259	261	515
17	909	909	930	910	1078	906	916	907	969	1147
18	407	407	416	407	419	407	408	411	419	440
19	424	418	425	420	427	425	448	447	446	487
20	600	591	592	604	627	618	673	707	694	1118
21	530	546	536	547	570	535	553	552	564	569
22	250	266	264	270	326	274	277	270	260	600
23	230	226	265	229	294	241	241	236	242	335
24	411	412	398	420	414	428	389	385	398	379
25	425	402	430	421	479	422	433	423	432	532
26	380	391	407	422	434	401	411	416	411	476
27	436	438	420	379	500	391	464	445	461	587
28	626	624	624	633	603	632	628	639	627	822
29	282	283	282	283	318	283	283	301	297	410
30	355	358	408	367	368	372	352	361	352	465

Table 15.3.3. RMSE of one-step MMSE forecasts of logged series (number of times each method has indicated rank).

Rank	PAR/	PAR/1	PAR/	PAR/	SUBSET/	SUBSET/	DSM	DES	SARIMA	MEANS
	PACF		AIC	BIC	AIC	BIC				
1	4	3	7	3	1	4	1	1	3	3
2	3	5	5	2	0	5	4	3	3	0
3	10	2	3	4	1	4	2	2	2	0
4	3	11	0	7	0	2	2	3	2	0
5	5	3	5	6	1	3	3	3	1	0
6	3	3	2	2	5	1	7	6	1	0
7	1	1	2	4	4	3	4	3	7	1
8	1	2	4	1	3	5	5	5	4	0
9	0	0	2	1	11	1	2	4	7	2
10	0	0	0	0	4	2	0	0	0	24
Rank-sum	110	119	127	134	230	145	166	173	178	268

Table 15.3.4. MAD of one-step MMSE forecasts of logged series (number of times each method has indicated rank).

Rank	PAR/	PAR/1	PAR/	PAR/	SUBSET/	SUBSET/	DSM	DES	SARIMA	MEANS
	PACF		AIC	BIC	AIC	BIC				
1	4	4	4	1	1	5	1	2	5	3
2	6	4	4	4	1	4	3	3	1	0
3	5	8	5	3	1	4	2	2	0	0
4	6	6	2	8	1	4	1	1	1	0
5	6	2	4	6	0	2	5	1	4	0
6	2	3	3	2	6	3	4	4	2	1
7	0	1	2	5	3	3	6	6	4	0
8	1	2	4	1	3	2	6	3	8	0
9	0	0	2	0	10	3	1	8	5	1
10	0	0	0	0	4	0	1	0	0	25
Rank-sum	105	111	137	135	221	133	175	185	180	268

Table 15.3.5. MAPE of one-step MMSE forecasts of logged series (number of times each method has indicated rank).

Rank	PAR/ PACF	PAR/1	PAR/ AIC	PAR/ BIC	SUBSET/ AIC	SUBSET/ BIC	DSM	DES	SARIMA	MEANS
1	3	5	3	1	3	5	1	1	5	3
2	5	4	3	5	2	3	2	5	1	0
3	4	7	4	4	1	3	5	1	0	1
4	7	2	5	7	0	3	3	1	2	0
5	7	6	4	2	1	3	1	2	4	0
6	2	2	1	5	1	4	6	5	4	0
7	1	1	2	5	2	3	5	7	3	1
8	1	3	6	1	3	2	6	2	4	2
9	0	0	2	0	11	4	0	6	7	0
10	0	0	0	0	6	0	1	0	0	23
Rank-sum	115	115	147	134	218	144	166	177	175	259

Table 15.3.6. MEDIAN APE of one-step MMSE forecasts of logged series (number of times each method has indicated rank).

Rank	PAR/ PACF	PAR/1	PAR/ AIC	PAR/ BIC	SUBSET/ AIC	SUBSET/ BIC	DSM	DES	SARIMA	MEANS
1	5	1	3	1	6	4	2	1	3	4
2	3	3	5	4	4	2	1	3	4	1
3	4	5	6	2	0	3	3	4	2	1
4	6	4	3	6	2	3	2	3	0	1
5	4	5	5	6	2	1	3	2	2	0
6	3	3	1	3	4	6	3	2	4	1
7	3	6	2	2	2	6	2	5	2	0
8	1	1	2	3	3	2	7	4	5	2
9	1	1	3	1	4	2	4	5	8	1
10	0	1	0	2	3	1	3	1	0	19
Rank-sum	123	150	131	154	160	156	190	175	177	234

The rank-sums for the models are the sums of the product of the rank and the associated table entry. Thus, models with lower rank-sums perform better than those with larger rank-sums. The models PAR/PACF, PAR/1, PAR/AIC, PAR/BIC, and SUBSET/BIC fare very well on the basis of all performance criteria. As expected, using the MEANS proves unsatisfactory in most cases. The MEANS has the worst overall performance and produces the largest RMSE for twenty-four of the series. Table 15.3.2 shows that in the three cases (rivers 2, 5, and 24) where the MEANS has the smallest RMSE there is very little difference between any of the forecasting methods. Moreover, in these three cases all methods have low MAPEs and RMSEs. At the other extreme, the best alternative to MEANS for rivers 7, 16, and 22 has a RMSE less than half that of MEANS. Next to the PAR models mentioned above, the DSM, DES, and SARIMA models perform about equally as well. The SUBSET/AIC model performance is disappointing, although not totally surprising. The large number of parameters associated with the SUBSET/AIC model does not provide a sufficiently parsimonious and flexible model for producing accurate forecasts. The importance of parsimony in forecasting models is discussed by Ledolter and Abraham (1981).

For several of the rivers, there are large discrepancies between the MAPE and MEDIAN APE criteria. This is found to be due to a defect in the absolute percentage error when the observed value is small. For example, the observed logged flow for river 14 for November, 1959, is 0.0024 and the PAR/PACF forecast is -0.746. This creates an absolute percentage error of over 31,000!.

The forecasting results reported thus far are for the logarithmic flows. To compare results in the untransformed domain, one converts the forecasts using [15.2.21]. Table 15.3.7 shows the forecasting findings for the RMSE of one-step ahead MMSE forecasts for the untransformed time series. Once again, the same PAR models perform the best. However, there are some differences in the untransformed and transformed forecasting results. In particular, notice the improvement of the MEANS model and the poor performance of the DES model. The DSM and SARIMA models still perform reasonably well and the SUBSET/AIC improves slightly.

Table 15.3.7. RMSE of one-step MMSE forecasts of the flows (number of times each method has indicated rank).

Rank	PAR/ PACF	PAR/1	PAR/ AIC	PAR/ BIC	SUBSET/ AIC	SUBSET/ BIC	DSM	DES	SARIMA	MEANS
1	2	5	7	0	4	4	3	0	2	3
2	5	4	5	6	0	3	3	0	4	0
3	11	3	3	5	0	3	2	0	1	2
4	6	6	4	5	1	5	1	1	0	1
5	1	8	2	8	2	6	0	0	3	0
6	4	3	3	3	5	1	6	0	1	4
7	1	0	4	1	2	5	3	5	4	5
8	0	1	1	1	5	2	7	5	4	4
9	0	0	1	1	10	1	4	1	8	4
10	0	0	0	0	1	0	1	18	3	7
Rank-sum	105	112	115	129	202	135	178	268	196	210

The *Wilcoxon signed rank test* (Wilcoxon, 1945) for paired data is used to test for statistically significant differences in the forecasting ability of the various procedures. In this test, which is also described in Appendix A23.2, the differences in the squares of the logarithmic forecast errors are computed. These differences are ranked in ascending order, without regard to sign, and assigned ranks from one to thirty-six. The sum of the ranks of all positive differences are then computed as T in [A23.2.3] and compared to tabulated values in order to ascertain if the forecasts from one model are significantly better than the forecasts from a competing model. These results are then used to examine the performance of the models across all thirty series. In this test, the $P-$ value associated with each T value is calculated by estimating the area in the tail of the distribution. Then, the Fisher (1970, p. 99) method for *combining significance levels* for one-sided tests is

$$-2\sum_{i=1}^{k}\ln(p_i) \approx \chi^2_{2k}, \qquad [15.3.1]$$

where p is the calculated P-value associated with T and k is the number of series considered in the test. This combination technique generally has greater power than alternative methods such as simply summing the T's.

Fisher's test is employed to compare the overall performance of the PAR/PACF model to that of the other competing models. In addition, the PAR/1 parameters are also estimated using the Yule-Walker equations to provide an additional model for comparison (PAR/YW1). In this way, identical forecasts produced by the PAR/PACF and PAR/YW1 models could be ignored, ensuring that only the differences in the forecasting procedures are compared. The results of Fisher's test are presented in Table 15.3.8. The PAR/PACF model is significantly better than all of the models except the PAR/1 and the PAR/AIC at the five-percent level. Since different estimation procedures are employed for the PAR/PACF and PAR/1 models, there are several forecasts that are almost, but not quite, identical. These are all included in the analysis, thus masking the differences in the performance of the two models. The PAR/YW1 model, however, employs the same estimation procedure, thus resulting in identical forecasts when an AR(1) model is identified for a particular month for the PAR/PACF model. This allows ties to be dropped from consideration, and results in the testing of only the differences between the two models. All series with fewer than five untied forecasts are dropped from consideration in this test. The results of this comparison indicates that when ties are ignored, the PAR/PACF model is better than the PAR/YW1 model at the two-percent level of significance. Although the PAR/AIC compares quite favourably with the PAR/PACF when the significance levels are combined, detailed examination of the results reveal that for three rivers the PAR/PACF forecasts significantly better at the five-percent level than the PAR/AIC. However, in no case are the PAR/AIC forecasts significantly superior to those of the PAR/PACF. Additional details are given in the thesis of Noakes (1984).

15.4 FORECASTING QUARTER-MONTHLY AND MONTHLY RIVERFLOWS

15.4.1 Introduction

The results of the forecasting study of Section 15.3 indicate that certain types of PAR models work better than other competing seasonal models when forecasting average monthly riverflow time series. In particular, PAR models identified using the sample periodic PACF

Table 15.3.8. Results of Fisher's test for the Wilcoxon tests when each model is compared to the PAR/PACF model.

Model	PAR/1	PAR/	PAR/	PAR/	SUBSET/	SUBSET/	ARMA/	ARMA/	SARIMA	MEANS
		YW1	AIC	BIC	AIC	BIC	DSM	DES		
χ^2	65.2	60.5	64.8	57.8	116.0	87.8	99.4	101.2	113.3	276
DF	60	40	58	40	60	60	60	60	60	60
SL (%)	30	2	25	3	10^{-4}	1	0.1	0.05	10^{-5}	10^{-13}

provide the most accurate forecasts. Because PPAR models are not used in the forecasting experiments of the previous section, one of the objectives of the forecasting study presented in this section as well as by Thompstone (1983) and Thompstone et al. (1985) is to show a forecasting study involving PPAR models, as well as other types of seasonal models. A second goal is to perform forecasting experiments with both quarter-monthly and monthly time series.

15.4.2 Time Series

The data sets used in this forecasting study are identical to those utilized in the seasonal modelling applications of Section 14.6 and the simulation experiments with seasonal models in Section 14.8.2. In particular, the time series consist of both the quarter-monthly and monthly flows of the rivers called the Alcan system, Rio Grande and Saugeen. For all of these series, the last three years on record are not used when the seasonal models given in Table 14.6.3 were fitted to the six series.

15.4.3 Seasonal Models

The seasonal models used in the forecasting experiments are those listed in Table 14.6.3. The seasonal models consist of the SARIMA (Chapter 12), deseasonalized (Chapter 13), PAR (Sections 14.2.2 and 14.3) and three types of PPAR (Section 14.5) models. The deseasonalized model called DES refers to the situation when the most appropriate ARMA model is fitted to a series fully deseasonalized using [13.2.3], for which the seasonal means and standard deviations are estimated by utilizing [13.2.4] and [13.2.5], respectively. In all cases, the logarithmic series are used and the fitted models are identical to those described in Section 14.6.

15.4.4 Forecasting Experiments

For each time series and fitted model given in Table 14.6.3, one step ahead forecasts are calculated for the additional three year period in each series. Besides these models, forecast errors are also calculated for a model labelled MEANS, which simply entails using the seasonal means of the logarithmic series as the one step ahead forecasts.

For each fitted model and logarithmic time series, one can calculate the *MSE (mean squared error)* of the forecast errors. In Table 15.4.1, the models are ranked according to their MSE's for each of the six time series. The lowest value of MSE is ranked as 1 whereas the highest number refers to the model which produces the least accurate forecasts for the series. Because the forecast errors are approximately normally distributed, they can be employed in the Pitman test (Pitman, 1939) described in Section 8.3 to determine if there are significant differences in the MSE's of the forecasts between any two seasonal models. Some representative results using the Pitman test are given in Table 15.4.2.

Table 15.4.1. MSE's of the one step ahead forecasts and ranking of the models according to MSE's.

Model	Quarter-monthly Series			Monthly Series		
	Alcan System	Rio Grande	Saugeen	Alcan System	Rio Grande	Saugeen
MEANS	0.4314 (7)	0.2831 (7)	0.6177 (7)	0.3476 (4)	0.3359 (7)	0.5712 (7)
SARIMA	0.2871 (6)	0.2293 (6)	0.4274 (6)	0.3325 (2)	0.2399 (6)	0.4883 (6)
DES	0.2634 (5)	0.2222 (3)	0.4117 (5)	0.3011 (1)	0.2381 (4)	0.4788 (5)
PAR	0.2575 (3)	0.2213 (1)	0.4070 (1)	0.3457 (3)	0.2301 (2)	0.4189 (2)
PPAR/50	0.2567 (2)	0.2235 (5)	0.4075 (2)		0.2289 (1)	0.4187 (1)
PPAR/20	0.2583 (4)	0.2221 (2)	0.4096 (3)		0.2321 (3)	0.4265 (4)
PPAR/05	0.2552 (1)	0.2227 (4)	0.4098 (4)		0.2381 (4)	0.4217 (3)

Note: The parenthetical figure ranks the MSE's for a given series from the lowest (1) to the highest (7).

Table 15.4.2. Pitman's correlation test statistics for comparing MSE's of one step ahead forecasts for seasonal models fitted to the quarter-monthly Saugeen riverflows.

	MEANS	SARIMA	DES	PAR	PPAR/50	PPAR/20	PPAR/05
MEANS		0.4476(-)	0.4735(-)	0.4814(-)	0.4831(-)	0.4799(-)	0.4792(-)
SARIMA*	0.4476(+)		0.1460(=)	0.1358(=)	0.1320(=)	0.1205(=)	0.1239(=)
ARMA/DES*	0.4735(+)	0.1460(=)		0.0504(=)	0.0424(=)	0.0201(=)	0.0195(=)
PAR*	0.4814(+)	0.1358(=)	0.0504(=)		0.0227(=)	0.0633(=)	0.0657(=)
PPAR/50*	0.4831(+)	0.1320(=)	0.0424(=)	0.0227(=)		0.0625*(=)	0.0524(=)
PPAR/20*	0.4799(+)	0.1205(=)	0.0201(=)	0.0633(=)	0.0625(=)		0.0024(=)
PPAR/05*	0.4792(+)	0.1239(=)	0.0195(=)	0.0657(=)	0.0524(=)	0.0024(=)	

(1) Table shows $|r|$ for Pitman's correlation test statistic.

(2) Difference in MSE's of forecasts is significant at 5% level if $|r| > 0.163$.

(3) A parenthetical = indicates the difference is not significant, a + indicates the row model is "better" than the column model (significant difference and smaller MSE), and a - indicates the row model is "worse" than the column model.

(4) * indicates the model is better or equal to all other models.

Consider the results for the MSE's in Table 15.4.1 for comparing the forecasting capabilities of the seasonal models. The simplistic MEANS model consistently provides the worst forecasts, and this confirms that the methods of time series analysis provide meaningful improvements in forecasting ability. In five of the six cases, SARIMA models provide the second largest forecast errors. This could lead to some doubts regarding the appropriateness of SARIMA models for forecasting the inflow series considered herein. As noted in Chapter 12 and elsewhere in Part VI, from a physical viewpoint SARIMA models are not well designed for modelling riverflow time series, like the one in Figure VI.1, because they cannot explicitly model stationarity within each season as well as a seasonally varying correlation structure.

In three cases (one quarter-monthly series and two monthly series), PPAR models provide the smallest MSE's of forecasts, while in two other cases (both quarter-monthly), PAR models produce the best forecasts. More generally, one sees that in five of the six cases, the smallest and second smallest MSE's are furnished by PAR or PPAR models; in four of the six cases, the four smallest MSE's are provided by PAR or PPAR models. All this suggests that PAR and PPAR models have appealing forecasting abilities for the series considered herein. In only one case, the Alcan system monthly inflow series, the DES provides the smallest MSE.

Table 15.4.2 shows the results of Pitman's correlation test for the case of the quarter-monthly flows of the Saugeen River. The statistic, $|r|$, for comparing MSE's between one step ahead forecasts for two seasonal models is described in Section 8.3.2. A parenthetical equal sign, (=), indicates that, at the 5% level, the difference between the row model errors and the column model errors is not significant. A parenthetical plus sign, (+), indicates that the row model provides significantly better forecasts than the column model, and a parenthetical negative sign, (-), indicates the contrary. An asterisk beside the label of the row model indicates that it provides forecasts which are, at the 5% level, equal to or better than forecasts from all other models considered. As is the situation in Table 15.4.2 and the results for the other 5 series which are not shown, in no case does a model furnish forecasts which are significantly better than forecasts from all other models.

In all three cases of quarter-monthly series, the DES, PAR, and PPAR/05 models give forecast errors which were statistically equivalent to or better than all other models. In two cases out of three, the SARIMA, PPAR/50 and PPAR/20 models are equal to or better than all other models with respect to their forecasting abilities. Only the MEANS model is, in all three cases, significantly worse than all other models.

Forecasting results for the monthly Alcan system inflows are inconclusive. No PPAR models are identified for this series, and there is no statistically significant difference in forecasts from the four other models. For the two other monthly series, forecasts are, at the 5% level of significance, indistinguishable for the SARIMA, DES, PAR, PPAR/50, PPAR/20 and PPAR/05 models. In the case of the monthly Rio Grande flows, the MEANS model is significantly worse than all other models, while for the monthly Saugeen riverflows, the MEANS model is indistinguishable from the SARIMA and DES models, but significantly worse than the others.

In summary, from the results of the Pitman test, it is difficult to conclude that, amongst the SARIMA, DES, PAR, and various PPAR models, one type of model is particularly outstanding with respect to its forecasting ability for the time series considered herein. However, it is interesting to recall that in three of the five cases for which they are identified, the PPAR models provide the smallest MSE's of forecasts (see Table 15.4.1), and in the other two cases it is the

PAR models which produces the best forecasts.

15.5 COMBINING FORECASTS ACROSS MODELS

15.5.1 Motivation

The selection of the "best" forecasting procedure is certainly a hopeful result of any forecasting study. Invariably, however, no one method will produce optimum forecasts in all cases. The task then becomes one of selecting the most appropriate forecasting procedure based upon the available information.

An alternative approach is to combine the forecasts from two or more procedures in accordance to their relative performances. In this way, it is hoped that the strengths of each method might be exploited. The successes achieved by combining economic forecasts are documented in several studies (Armstrong and Lusk, 1983; Bates and Granger, 1969; Bordley, 1982; Granger and Ramanathan, 1984; Makridakis et al., 1982; Newbold and Granger, 1974; Winkler and Makridakis, 1982). Within the field of water resources, McLeod et al. (1986) present experimental results on combining hydrologic forecasts which are also described in Sections 15.5.3 and 18.4.2 of this book.

In the next subsection, techniques for combining forecasts are given. Subsequently, in Section 15.5.3, seasonal riverflow forecasts generated using both PAR and SARIMA models are combined in an attempt to achieve improved forecasts. Within Section 18.4.2, seasonal riverflow forecasts from TFN, PAR and conceptual or physically based models are optimally combined in forecasting experiments.

15.5.2 Formulae for Combining Forecasts

There are certainly countless ways of combining forecasts from different forecasting procedures to arrive at a combined forecast. The simplest is probably to weight each forecast equally. If there are k forecasts available, the combined forecast f_c, would be

$$f_c = \sum_{i=1}^{k} w_i f_i \qquad [15.5.1]$$

where f_i is the forecast produced by the ith model, w_i is the weighting factor for the ith forecast and $w_i = w_j = 1/k$ for all i and j.

It would be expected that a better combination of forecasts could be obtained if the statistical properties of the forecast errors were considered. Winkler and Makridakis (1983) point out that if the covariance matrix of the forecast errors from k methods, Σ, is known, then the optimal weights are given by

$$w_i = \frac{\sum_{j=1}^{k} \alpha_{ij}}{\sum_{h=1}^{k}\sum_{j=1}^{k} \alpha_{hj}} \qquad [15.5.2]$$

where the α_{ij} terms are the elements of Σ^{-1}. In practice, Σ is not known and must be estimated. Estimates of the weights in [15.5.2] can be calculated from the inverse of $\hat{\Sigma}$ where

$$\hat{\Sigma}_{ij} = \mathrm{v}^{-1} \sum_{h=t-\mathrm{v}}^{t-1} e_h^{(i)} e_h^{(j)} \qquad [15.5.3]$$

$e_t^{(i)}$ is the percentage error for method i at time t and v is the number of previous forecast errors employed to calculate w_i.

In the study concerning the combination of economic forecasts by Winkler and Makridakis (1983), these authors found that estimating Σ^{-1} and calculating the weights using [15.5.2] gave the poorest results. One of the preferred procedures in their study was to ignore the correlation between the forecast errors. In this case, the forecast weights were calculated as

$$\hat{w}_i = \frac{\left[\sum_{h=t-\mathrm{v}}^{t-1} e_h^{(i)2}\right]^{-1}}{\sum_{j=1}^{p}\left[\sum_{h=t-\mathrm{v}}^{t-1} e_h^{(j)2}\right]^{-1}} \qquad [15.5.4]$$

where $e_t^{(i)}$ and v are as defined previously. This approach ensures that all of the estimated weights are greater than or equal to zero.

An alternative approach to calculating the combining weights when seasonal data are considered is developed by McLeod et al. (1986). In this procedure, the model residuals are employed to calculate the residual variance for each season. If two forecasts are to be combined, then the weights are calculated for each season such that

$$w_{1,j} = \frac{\sum_{k=1}^{n}[a_{j+(k-1)s}^{(1)}]^2}{\sum_{k=1}^{n}[a_{j+(k-1)s}^{(1)}]^2 + \sum_{k=1}^{n}[a_{j+(k-1)s}^{(2)}]^2} \qquad [15.5.5]$$

and

$$w_{2,j} = \frac{\sum_{k=1}^{n}[a_{j+(k-1)s}^{(2)}]^2}{\sum_{k=1}^{n}[a_{j+(k-1)s}^{(1)}]^2 + \sum_{k=1}^{n}[a_{j+(k-1)s}^{(2)}]^2} \qquad [15.5.6]$$

where $w_{1,j}$ is the weight assigned to forecasting procedure one for the jth season, $w_{2,j}$ is the weight assigned to forecasting procedure two for the jth season, $a_t^{(i)}$ is the residual at time t for the ith model, n is the number of years of data and s is the number of seasons per year. Since the data are seasonal, the forecast error variance might be expected to be seasonal and, hence, this procedure should account for this seasonality.

15.5.3 Combining Average Monthly Riverflow Forecasts

The thirty average monthly riverflow time series listed in Table 15.3.1 and referred to in Section 15.3.2 are the data sets employed in the experiments for combining forecasts among two models. As is also the situation in Section 15.3.3, the last three years or 36 observations are omitted from each of the data sets. Subsequently, after taking natural logarithms of each time series both PAR/PACF and SARIMA models are fitted to each of the truncated logarithmic sequences. Recall from Section 15.3.3 that PAR/PACF refers to a calibrated PAR model that is identified using the sample periodic PACF. The same 36 one-step-ahead forecasts calculated in Section 15.3.4 for each of these two models and each of the thirty series are employed in the combination experiments reported upon here.

The monthly logarithmic forecasts produced by the PAR/PACF and SARIMA models are combined using some of the procedures outlined previously in Section 15.5.2. Specifically, the combining weights are calculated using [15.5.4] with v = 3, 6, 9, and 12. In addition, seasonal combining weights are also determined employing [15.5.5] and [15.5.6]. The combined forecasts are then compared on the basis of MSE's. A summary of the results is presented in Table 15.5.1. The CMB-SEAS entries refer to the combined forecasts produced when separate weights are calculated for each season. The CMB-v entries represent the combined forecasts when the previous v forecast errors are employed to calculate the combining weights. The results show that, in general, the combined forecasts do not constitute an improvement over the PAR/PACF forecasts, regardless of the procedure utilized to calculate the combining weights. This is because the PAR family of models has a better mathematical design for forecasting an average monthly riverflow series like the one in Figure VI.1 while the SARIMA model is more suitable for forecasting series such as those in Figures VI.2 and VI.3. Accordingly, the PAR model forecasts better than the SARIMA model and attempting to combine inferior forecasts with better ones does not improve the situation for the PAR forecasts. Conversely, the SARIMA forecasts are almost always improved by combining them with PAR/PACF forecasts. Finally, a comparison of the various procedures for combining the forecasts seems to indicate that the more information employed to estimate the combining weights the better the forecasts.

Table 15.5.1. Percentage of times model A gives better values for forecasting a series than model B.

Model A	Model B						
(1)	PAR/PACF (%) (2)	SARIMA (%) (3)	CMB-SEAS (%) (4)	CMB-3 (%) (5)	CMB-6 (%) (6)	CMB-9 (%) (7)	CMB-12 (%) (8)
PAR/PACF	0	70	56.7	60	60	56.7	56.7
SARIMA	30	0	20	16.7	20	20	20
CMB-SEAS	43.4	80	0	46.7	43.3	56.7	56.7
CMB-3	40	83.3	53.3	0	50	46.7	33.3
CMB-6	40	80	46.7	50	0	43.3	26.7
CMB-9	43.3	80	43.3	53.3	56.7	0	30
CMB-12	43.3	80	43.3	66.7	73.3	70	0

15.6 AGGREGATION OF FORECASTS

Suppose that one wishes to forecast future average annual riverflows for a given river for which the average monthly and hence also the annual values are known. One approach is to fit a nonseasonal time series model such as an ARMA model to the yearly data and then employ this model for forecasting annual values. Another procedure is to fit an appropriate seasonal model like the PAR model of Sections 14.2.2 and 14.3 to the average monthly series and then utilize this model to forecast the next 12 months. The sum of these 12 monthly forecasts would represent an *aggregated forecast* for the yearly value.

Noakes (1984, Ch. 6) carried out forecasting experiments with riverflow time series to ascertain if aggregated forecasts can improve the accuracy of forecasts determined for a larger time interval. For various yearly and seasonal time series models, Noakes found that for the data sets that he considered, the aggregated forecasts for annual values were generally not as good as those produced by the annual models.

For further research on aggregation of forecasts the reader can refer to Tiao (1972) and Tiao and Wei (1976). Moreover, a discussion on disaggregation and aggregation in time series modelling within the hydrological literature is given in Section 20.5.2.

15.7 CONCLUSIONS

As explained in Section 15.2, MMSE forecasts can be easily calculated for all the seasonal models presented in Part VI. The results of the forecasting experiments of Section 15.3 for 30 monthly riverflow series clearly indicate that PAR models identified using the sample periodic PACF forecast significantly better than SARIMA, deseasonalized, and PAR models identified using techniques other than the sample periodic PACF. When PPAR models are also considered, the forecasting studies of Section 15.4 show that PPAR models also forecast quite well. Finally, forecasts can be combined across models in an attempt to achieve improved forecasts by using procedures described in Section 15.5.2 and applied to monthly riverflow time series in Section 15.5.3.

PROBLEMS

15.1 Select a seasonal time series for which you think may be appropriate to fit a SARIMA model and then carry out the following tasks:

(a) Examine suitable exploratory data analysis graphs for discovering the key statistical characteristics of the time series.

(b) Remove the final year of observations from the time series and then by following the three stages of model construction, fit a SARIMA model to the remaining values.

(c) Calculate the MMSE forecasts and 90% probability limits for the last year of observations to which the model was not fitted. Clearly explain how you perform your calculations.

(d) Plot the MMSE forecasts and 90% probability limits on a graph with the historical observations for the final year. Determine the accuracy of the forecasts and comment upon any interesting findings.

15.2 Follow the instructions in problem 15.1 for a deseasonalized model.

15.3 Carry out the instructions in problem 15.1 for a PAR model.

15.4 Choose a time series to which it seems reasonable to fit SARIMA, deseasonalized and PAR models. For each of these models follow the instructions of question 15.1. Additionally, for the time series under study, compare the forecasting capabilities of the three seasonal models and ascertain if one model forecasts significantly better than another.

15.5 Makridakis et al. (1982) carry out forecasting experiments for a range of models fitted to 1001 time series consisting of yearly, monthly and quarter-monthly economic data sets. After reading their paper, respond to the following questions:

(a) Outline the major findings of their study.

(b) Describe the main steps these authors followed in executing their forecasting experiments and comparing the forecasting results for the various models and data sets.

(c) Explain the commonalities and differences between the procedures used by the authors of the forecasing paper for carrying out their forecasting experiments with those employed in this book.

15.6 Carry out the instructions of the previous question for the paper by Newbold and Granger (1974).

15.7 From your field of study, pick out a set of three or more seasonal time series that are of direct interest to you. After fitting appropriate time series models from Part VI to the first portion of each series, execute forecasting experiments to ascertain which class or classes of models provide the most accurate forecasts. A summary of how to perform a systematic forecasting study is given in Figure 8.3.1.

15.8 Employing procedures described in Section 15.5.2, combine forecasts among pairs of models used in problem 15.7 in order to ascertain if enhanced forecasts can be found. Comment upon any interesting discoveries that you may make.

15.9 Summarize the main research findings of Tiao (1972) as well as Tiao and Wei (1976) on the aggregation of forecasts.

15.10 The aggregation of forecasts is discussed in Section 15.6. Select an average monthly riverflow time series and then do the following:

(a) Fit a PAR model to all but the last three years of the monthly series. Employ this model to forecast the last 36 values. For each of the last 3 years, determine the aggregated forecast for each year.

(b) Fit an ARMA model to the average annual series for which the last 3 years are left out. Employ this calibrated model to forecast the next three years.

(c) Compare the accuracy of the annual forecasts obtained in points (a) and (b) and comment upon the results.

(d) Discuss the annual forecasting results when only one step ahead forecasts are employed in parts (a) and (b).

REFERENCES

In addition to the references listed here, the reader may also wish to refer to forecasting references given at the ends of Chapters 1, 8 and 18.

AGGREGATION OF FORECASTS

Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregation of time series. *Biometrika*, 59:525-531.

Tiao, G. C. and Wei, W. S. (1976). Effect of temporal aggregation on the dynamic relationship of two time series. *Biometrika*, 63:513-523.

COMBINING FORECASTS

Armstrong, J. S. and Lusk, E. J. (1983). Commentary on Makridakis time series competition (m-competition). *Journal of Forecasting*, 2:259-311.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly* (Journal of the Operational Research Society), 10:451-468.

Bordley, R. F. (1982). The combination of forecasts - a Bayesian approach. *Journal of the Operational Research Society*, 33:171-174.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3:197-204.

McLeod, A. I., Noakes, D. J., Hipel, K. W. and Thompstone, R. M. (1987). Combining hydrological forecasts. *Journal of Water Resources Planning and Management*, American Society of Civil Engineers, 113(1):29-41.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A*, 137:131-165.

Winkler, R. L. and Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A*, 146:150-157.

FORECASTING

Brocha, M. (1978). Computerized rivers. *Science Dimension*, 10:18-20.

Camacho, F. (1990). Forecasts for Niagara River flows at Ontario Hydro. *Proceedings of the Great Lakes Water Level Forecasting and Statistics Symposium*, held May 17-18, 1990, in Windsor, Ontario, Canada, 73-80.

Delleur, J. W., Tao, P. C. and Kavvas, M. L. (1976). An evaluation of the practicality and complexity of some rainfall and runoff time series models. *Water Resources Research*, 12(5):953-970.

Granger, C. W. J. and Newbold, P. (1976). Forecasting untransformed series. *Journal of the Royal Statistical Society, Series B*, 38(2):189-203.

International Association of Hydrological Sciences (1980). *Hydrological Forecasting Symposium, Proceedings of the Oxford Hydrological Forecasting Symposium*, held April 15-18, 1980. IAHS-AISH publication No. 129.

Ledolter, J. and Abraham, B. (1981). Parsimony and its importance in time series forecasting. *Technometrics*, 23(4):411-414.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, 1:111-153.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A*, 137:131-165.

Noakes, D. J. (1984). Applied Time Series Modelling and Forecasting. PhD thesis, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Noakes, D. J., McLeod, A. I., and Hipel, K. W. (1983). Forecasting experiments with seasonal hydrologic time series models. Technical Report 117-XM-220283, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Noakes, D. J., McLeod, A. I., and Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, 1:179-190.

Thomas, H. A. and Fiering, M. B. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Maas, A., Hufshmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A. and Fair, M. G., Editors, *Design of Water Resources Systems*, pages 459-493. Harvard University Press, Cambridge, MA.

Thompstone, R. M. (1983). *Topics in Hydrological Time Series Modelling*. PhD thesis, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1983). Transfer function-noise modelling for powerhouse inflow forecasting. *INFOR*, 21:259-269.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985). Forecasting quarter-monthly riverflow. *Water Resources Bulletin* 21(5):731-744.

STATISTICS AND TIME SERIES ANALYSIS

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Revised Edition.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, England.

McClave, J. T. (1975). Subset autoregression. *Technometrics*, 17(2):213-220.

Morgan, J. A. and Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14(2):317-325.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31:9-12.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80-83.

PART VII

MULTIPLE INPUT - SINGLE OUTPUT MODELS

In many environmental systems, a single output or response variable is "caused" by one or more input or covariate series. For example, riverflows are caused by physical variables such as precipitation and temperature. To formally model the dynamic relationships which exist between a single output variable and the multiple input variables, a **transfer function-noise (TFN) model** can be employed. Qualitatively, a TFN model can be written as

single output = dynamic component + noise

where the dynamic component models the manner in which each input or covariate series affects the dynamic response of the output and the noise accounts for the stochastic disturbance in the system which cannot be modelled by the dynamic component. Because the behaviour of the output is dependent upon the way the input series affect the output over time, the overall TFN model is often referred to as a **dynamic model.**

An array of useful tools are available for **constructing TFN models** when following the identification, estimation and diagnostic check stages of model development. At the **identification** stage, a transfer function can be designed for mathematically describing the dynamic relationship over time which exists between each input and the output. An appropriate ARMA or ARIMA model can be identified as the autocorrelated noise component in the overall TFN model. Following the **estimation** of the model parameters and **checking** that the fitted model adequately describes the dynamic system being modelled, the calibrated TFN model can be used for applications such as forecasting and simulation. As is demonstrated in Part VII, the presence of the input variables in the model allows for a more accurate description of the physical system which in turn means more accurate forecasts (Chapter 18) and realistic simulated values can be produced by the model. Furthermore, TFN models can be built for either seasonal or nonseasonal time series for which the data points are evenly spaced over time.

In certain situations it may not be obvious if one physical variable causes another. For instance, do sunspot numbers cause riverflows? Consequently, in Chapter 16 statistical procedures are presented as **exploratory data analysis** tools for investigating possible **causal relationships** between two variables. When meaningful relationships are detected between two series using what is called the **residual cross-correlation function,** a TFN model can be constructed as a **confirmatory data analysis** procedure for rigorously describing the mathematical relationship between the input and output. In Chapter 17, comprehensive methods for constructing TFN models with a single output and multiple inputs are explained for both seasonal and nonseasonal time series using a number of interesting hydrological applications. Subsequent to calibrating a TFN model, the fitted model can be employed for **forecasting** by following the procedures of Chapter 18.

Sometimes the dynamic characteristics of a system may be changed by the imposition of one or more external interventions. For example, in environmental engineering, pollution abatement facilities are built to reduce the levels of certain pollutants. The stochastic effects upon the mean level of the output can be rigorously modelled using **intervention analysis.** As will be thoroughly explained in Chapters 19 and 22, the intervention model is in fact a special type of TFN model. An extensive description of exploratory and confirmatory data analysis procedures for use in intervention analysis is presented in these chapters. Subsequent to calibrating a TFN model, the fitted model can be employed for **forecasting** by following the procedures of Chapter 18.

CHAPTER 16

CAUSALITY

16.1 INTRODUCTION

Is it possible to substantiate the claim of a Soviet hydrologist that yearly sunspot numbers have a significant affect upon the annual flows of the Volga River? What is the influence of temperature upon the price of wheat? In other words, how and when can one say that one phenomenon definitely causes another?

The foregoing kinds of questions have been baffling scientists for decades and previously some research had been carried out in an attempt to answer them. For example, Brillinger (1969) and Rodriguez-Iturbe and Yevjevich (1968) employed cross-spectral and other statistical methods to investigate relationships between natural time series. However, comprehensive statistical tools are now available to assist in solving causality problems and these useful techniques have yet to be applied to a large variety of environmental data sets. Consequently, the purpose of this chapter is to present flexible statistical procedures for formally answering causality questions and then to apply the methodologies to a wide range of natural time series. In particular, Granger's (1969) definition of *causality* is first defined and then it is explained how a *cross-correlation analysis of the residuals* from the stochastic models fitted to two series, can be employed to detect causal relationships (Pierce and Haugh, 1977). In the section on *applications*, a large number of interesting cross-correlation studies are carried out to detect possible causal relationships between many different phenomena. The time series studied include sunspot numbers, annual and monthly temperatures, seven annual riverflow series, Beveridge wheat price indices, and tree ring widths. Contrary to the suggestion of Smirnov (1969), it is found that annual sunspot numbers do not significantly affect the yearly flows of the Volga River in Russia. Other causality studies demonstrate that temperatures for certain months of the year can significantly affect the annual flows of rivers and also the price of wheat.

Upon detecting significant causal connections between two phenomena, the information from the cross-correlation analysis can be used to design a *transfer function-noise (TFN) model* to describe explicitly the mathematical relationship between the two data sets (Haugh and Box, 1977; Box and Jenkins, 1976, Ch. 11). In Chapter 17, the *construction* of TFN models which can handle a single output series and one or more input series, is thoroughly explained for the identification, estimation and diagnostic check stages of model development. Moreover, in Chapter 18, it is explained how one can calculate optimal *forecasts* using a TFN model. As would be expected, the information contained in the input or covariate series in a TFN model allows one to obtain more accurate forecasts for the output series. Finally, for the original presentation of the main contents of Chapter 16, the reader can refer to the paper of Hipel et al. (1985).

16.2 CAUSALITY

16.2.1 Definition

Wiener (1956) originally formulated a definition of causality between two time series, which is suitable for empirical detection and verification of meaningful relationships. More recently, Granger (1969) presented a formal definition of causality while Pierce and Haugh (1977) expanded upon the work of Granger (1969) and gave a comprehensive survey regarding research on causality in temporal systems. Other research which is related to Granger's (1969) definition of causality can be found by referring to the appropriate statistical literature (see for example Jenkins and Watts (1968), Haugh (1972, 1976), Haugh and Box (1977), and McLeod (1979)).

Granger (1969) defines *causality* between two time series in terms of predictability. A variable X causes another variable Y, with respect to a given universe or information set that includes X and Y, if the present Y can be better predicted by using past values of X than by not doing so (all other relevant information (including the past of Y) being used in either case). This definition of causality does not require the system to be linear but when it is, linear predictions are compared. To be more specific, let X_t and Y_t be two time series and let A_t for $t = 0,\pm 1 \pm 2, \ldots,$ be the given information set that includes at least X_t and Y_t. Allow $\bar{A}_t = \{A_s : s < t\}$, $\tilde{A}_t = \{A_s : s \le t\}$ and in a similar fashion define $\bar{X}_t$, $\tilde{X}_t$, $\bar{Y}_t$, and $\tilde{Y}_t$. Given the information set A_t, let $P_t(Y|A_t)$ be the minimum mean square error one step ahead predictor of Y_t and denote the resulting mean square error by $\sigma^2(Y|A_t)$. According to Granger (1969), X causes Y if

$$\sigma^2(Y|\bar{A}_t, \bar{X}_t) < \sigma^2(Y|\bar{A}_t) \qquad [16.2.1]$$

while X causes Y instantaneously if

$$\sigma^2(Y|\bar{A}_t, \tilde{X}_t) < \sigma^2(Y|\bar{A}_t) \qquad [16.2.2]$$

Causality from Y to X can be defined in the same way. *Feedback* occurs when X causes Y and Y also causes X.

16.2.2 Residual Cross-Correlation Function

To ascertain the type of causality relationship that exists between X and Y, the properties of the cross-correlations are examined for the prewhitened series. When *prewhitening* discrete time series such as X_t and Y_t, the first step is to consider suitable *transformations* to form the transformed series, x_t and y_t. The reasons for transforming the series include stabilizing the variance, improving the normality assumption, eliminating trends, removing seasonality, and getting rid of nonstationarity. The selected transformations should allow x and y to be related causally in the same manner as X and Y when considering Granger's (1969) definition of causality. In practice, causality is preserved by many of the common types of transformations. For example, often the given series may be transformed by the *Box-Cox transformation* (Box and Cox, 1964) given in [3.4.30] to remove non-normality and heteroscedasticity in the model residuals and following this the data may be differenced as in [4.3.3] to render the data stationary. As is explained in Section 13.2.2, when dealing with seasonal geophysical series the data may be transformed using a Box-Cox transformation and subsequent to this the seasonality may be removed by invoking an appropriate *deseasonalization* technique. For instance, when modelling an average monthly riverflow series, often the series is first transformed by taking natural logarithms and then each data point is deseasonalized by subtracting out the monthly mean and dividing this by the monthly standard deviation as in [13.2.3]. A Box-Cox transformation such as natural logarithms should not alter causality relationships for series consisting of all positive values, since the manner in which one series affects the predictability of another will not be changed by a strictly monotonic transformation that preserves the same relative position of every data point in the series. Deseasonalizing each time series is equivalent to removing a periodic component to eliminate seasonality where the periodic component is ultimately due to hydrologic factors such as precipitation and temperature. Because the deseasonalization parameters are estimated from the historical data and are assumed to be the same in the future, the deseasonalization should not alter the causality relationship existing in the original series when entertaining Granger causality. However, the periodic portion still constitutes one of the components needed to form the overall seasonal series.

The second step in the prewhitening procedure is to fit appropriate stochastic models to the x_t and y_t series in order to obtain white noise residuals. For instance, when the transformed series are nonseasonal, it may be suitable to fit the ARMA model in [3.4.4] to x_t and y_t such that

$$\phi_x(B)(x_t - \mu_x) = \theta_x(B)u_t \qquad [16.2.3]$$

and

$$\phi_y(B)(y_t - \mu_y) = \theta_y(B)v_t \qquad [16.2.4]$$

where μ_x is the theoretical mean of the x_t series; B is the backward shift operator defined by $Bx_t = x_{t-1}$ and $B^k x_t = x_{t-k}$ where k is a positive integer; $\phi_x(B) = 1 - \phi_{x,1}B - \phi_{x,2}B^2 - \cdots - \phi_{x,p_x}B^{p_x}$, is the nonseasonal AR operator of order p_x such that the roots of the characteristic equation $\phi_x(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the $\phi_{x,i}, i = 1,2, \cdots, p_x$, are the nonseasonal AR parameters; $\theta_x(B) = 1 - \theta_{x,1}B - \theta_{x,2}B^2 - \cdots - \theta_{x,q_x}B^{q_x}$, is the nonseasonal MA operator of order q_x such that the roots of $\theta_x(B) = 0$ lie outside the unit circle for invertibility and $\theta_{x,j} = 1,2, \cdots, q_x$, are the nonseasonal MA parameters; u_t is white noise (also called innovation or disturbance) that has a mean of zero and variance of σ_u^2 ; and similar definitions to μ_x, $\phi_x(B)$, $\theta_x(B)$, and u_t hold for μ_y, $\phi_y(B)$, $\theta_y(B)$, and v_t, respectively. As mentioned in Section 3.4.2, to indicate the orders of the AR and MA operators of the models in [16.2.3] or [16.2.4], the notation ARMA(p,q) is employed. Because of the linear nature of the operators in [16.2.3] and [16.2.4], this insures that u and v are causally related in the same way as X and Y. Of course, if the data were seasonal an appropriate seasonal model, such as one of those given in Chapters 12 to 15, could be used to prewhiten each series.

Subsequent to prewhitening of the time series, the *cross-correlation function (CCF)*, at lag k between the u_t and v_t series in [16.2.3] and [16.2.4], respectively, can be considered using

$$\rho_{uv}(k) = E[u_t v_{t+k}] / (E[u_t^2]E[v_t^2])^{1/2} \qquad [16.2.5]$$

Due to the form of [16.2.5], the values of the CCF can range from negative one to positive one. Unlike the ACF, the CCF is not usually symmetric about lag zero and therefore the properties of $\rho_{uv}(k)$ must be examined for $k = 0, \pm 1, \pm 2, \cdots$. In addition to reflecting the type of linear dependence between u and v and consequently between X and Y, $\rho_{uv}(k)$ gives the kind of causality relationship between these variables for linear systems.

As explained by Pierce and Haugh (1977), there are many possible types of causal interactions between X and Y which can be characterized by the properties of $\rho_{uv}(k)$. Using the results of Pierce and Haugh (1977, p. 276, Table 3), some of the important causal relationships are categorized according to the restrictions on $\rho_{uv}(k)$ in Table 16.2.1. Due to the findings of Price (1979) and also Pierce and Haugh (1979), any of the relationships in Table 16.2.1 which involve instantaneous causality are only valid when there is no feedback. The entries in Table 16.2.1 are self explanatory. For example, when there is unidirectional causality from X to Y, $\rho_{uv}(k) \neq 0$ for the $k > 0$, $\rho_{uv}(k) = 0$ for all $k < 0$, and $\rho_{uv}(0)$ may either be zero or else have some real non-zero value. For the case where Y does not cause X at all, there is no instantaneous causality between X and Y since $\rho_{uv}(0) = 0$.

When there is *feedback* between two variables, one variable can cause the other and vice versa. Although feedback is not too common in many natural problems, in economics, for example, inflation can cause unemployment which in turn affects inflation. As indicated in Table 16.2.1, $\rho_{uv}(k)$ is nonzero at both positive and negative lags if there is feedback between X and Y.

When checking for the type of causality between two given time series the estimated CCF of the model residuals must be examined to ascertain which values are significantly different from zero. Suppose that two sequences x_t and y_t are given for $t = 1,2, \ldots, n$. By utilizing [16.2.3] and [16.2.4] or other appropriate linear models, the two series can be prewhitened to obtain the estimated innovation series or residuals, $\hat{u}_t$ and $\hat{v}_t$, respectively. The residual CCF at lag k between $\hat{u}_t$ and $\hat{v}_t$ is estimated using

$$r_{\hat{u}\hat{v}}(k) = c_{\hat{u}\hat{v}}(k) / [c_{\hat{u}}(0) c_{\hat{v}}(0)]^{1/2} \qquad [16.2.6]$$

where

$$c_{\hat{u}\hat{v}}(k) = \begin{cases} n^{-1}\sum_{t=1}^{n-k} \hat{u}_t \hat{v}_{t+k} & k \geq 0 \\ n^{-1}\sum_{t=1-k}^{n} \hat{u}_t \hat{v}_{t+k} & k < 0 \end{cases}$$

is the *estimated cross-covariance function* at lag k between the residual series; $c_{\hat{u}}(0) = n^{-1}\sum_{t=1}^{n}\hat{u}_t^2$ is the sample variance of the $\hat{u}_t$ sequence; and $c_{\hat{v}}(0) = n^{-1}\sum_{t=1}^{n}\hat{v}_t^2$ is the estimated variance of the $\hat{v}_t$ series.

Table 16.2.1. Causal relationships between two variables.

RELATIONSHIPS	RESTRICTIONS ON $\rho_{uv}(k)$
X causes Y	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$
Y causes X	$\rho_{uv}(k) \neq 0$ for *some* $k < 0$
Instantaneous Causality	$\rho_{uv}(0) \neq 0$
Feedback	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$ and for some $k < 0$
X causes Y but not instantaneously	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$ and $\rho_{uv}(0) = 0$
Y does not cause X	$\rho_{uv}(k) = 0$ for *all* $k < 0$
Y does not cause X at all	$\rho_{uv}(k) = 0$ for *all* $k \leq 0$
Unidirectional causality from X to Y	$\rho_{uv}(k) \neq 0$ for some $k > 0$ and $\rho_{uv}(k) = 0$ for either (a) all $k < 0$ or (b) all $k \leq 0$
X and Y are only related instantaneously	$\rho_{uv}(0) \neq 0$ and $\rho_{uv}(k) = 0$ for all $k \neq 0$
X and Y are independent	$\rho_{uv}(k) = 0$ for *all* k

The residual CCF can be plotted against lag k for $k \simeq -n/4$ to $k \simeq n/4$. In order to plot confidence limits, the distribution of the residual CCF must be known. Assuming that the x_t and y_t series are independent (so $\rho_{uv}(k) = 0$ for all k), Haugh (1972, 1976) shows that for large samples $\hat{r}_{\hat{u}\hat{v}}(k)$ is normally independently distributed with a mean of zero and variance of $1/n$. Consequently, to obtain the approximate 95% confidence limits a line equal to 1.96 $n^{-1/2}$ can be plotted above and below the zero level for the residual CCF. McLeod (1979) presents the asymptotic distribution of the residual CCF for the general case where the x_t and y_t series do not have to be independent of each other and, consequently, more accurate confidence limits can be obtained by utilizing his results.

One reason why the residual CCF is examined rather than the CCF for the x_t and y_t series, is that it is much easier to interpret the results from a plot of $r_{\hat{u}\hat{v}}(k)$. This is because when both the x_t and y_t series are autocorrelated, the estimates of the CCF for x_t and y_t can have high variances and the estimates at different lags can be highly correlated with one another (Bartlett, 1935). In other words, the distribution of the estimated CCF for x_t and y_t is more complex than the distribution of $r_{\hat{u}\hat{v}}(k)$. Monte Carlo studies executed by Stedinger (1981), demonstrate the advantages of prewhitening two series before calculating their CCF. Additionally, from an intuitive point of view it makes sense to examine the residual CCF. Certainly, if the *driving mechanisms* or residuals of two series are significantly correlated, then meaningful relationships would exist between the original series.

From an examination of the residual CCF, the type of relationship existing between X and Y can be ascertained by referring to the results in Table 16.2.1. Suppose, for example, the X variable is precipitation and the Y variable is riverflow. From a physical understanding of hydrology, it is obvious that precipitation causes riverflow. This knowledge would be mirrored in a plot of the residual CCF for these two series. For $k \geq 0$ there would be at least one value of $r_{\hat{u}\hat{v}}(k)$ which is significantly different from zero. However, all values of the residual CCF for $k < 0$ would not be significantly different from zero. In situations where the type of causality between two series is not known (for instance, do sunspots cause riverflows), an examination of the residual CCF can provide valuable insight into the problem (see Section 16.3).

Formal tests of significance may also be derived when examining causal relationships (see, for example, McLeod (1979) and Pierce (1977)). Suppose that it is known a priori that Y does not cause X so that $\rho_{uv}(k) = 0$ for $k < 0$ (for instance riverflows do not cause precipitation). Consequently, one may wish to test the null hypothesis that X does not cause Y and hence $\rho_{uv}(k) = 0$ for $k = 1,2,\ldots,L$, where L is a suitably chosen lag such that after L time periods it would be expected there would not be a relationship between the x_t and y_t series. The statistic

$$Q_L = n^2 \sum_{k=0}^{L} \frac{r_{uv}^2(k)}{n-k} \qquad [16.2.7]$$

is then approximately distributed as $\chi^2(L+1)$. A significantly large value for Q_L would mean that the hypothesis should be rejected and, therefore, X causes Y.

A limitation of the methods explained in this section is that they are only useful when describing the relationships between two time series. If three or more time series are mutually related, then analyzing them only two at a time may lead to finding spurious relationships among them. Consequently, further research on causality between linear systems is still required. Nevertheless, as shown by the applications in the next section, in many situations bivariate causality studies are of direct interest to the practitioner.

When sufficient data are available, an alternative approach for detecting causal linear relationships is to work in the frequency domain rather than the time domain by employing the *coherence function*. An advantage of this procedure is that it can be extended for handling multiple-input and multiple-output systems (Bendat and Piersol, 1980).

16.3 APPLICATIONS

16.3.1 Data

For a long time, hydrologists have been attempting to ascertain the impact of exogenous forces upon specific hydrological and meteorological phenomena. In many instances, the great complexity of the physical problem at hand has precluded the development of suitable physical or statistical models to describe realistically the situation. Consequently, a wide range of phenomena are now studied in order to detect and model meaningful dynamic relationships.

The time series investigated are listed in Table 16.3.1. Except for monthly temperatures from the English Midlands, all of the data sets consist of annual values. The sunspot numbers, annual and monthly temperatures, seven riverflow series in m^3/s where each average yearly flow is calculated for the water year from October 1st of one year to September 30th of the next year, and Beveridge wheat price indices, are obtained from articles by Waldemeier (1961), Manley (1953, pp. 255-260), Yevjevich (1963), and Beveridge (1921), respectively. The tree ring widths given in units of 0.01 mm are for Bristlecone Pine and were received directly from V.C. LaMarche of the Laboratory of Tree Ring Research, University of Arizona, Tuscon, Arizona. The length and accuracy of the tree ring series make it a valuable asset in cross-correlation studies for determining the effects of external variables such as temperature and the amount of sunlight. The reason for considering the Beveridge wheat price index data is that the series could be closely related to climatic conditions and, therefore, may be of interest to hydrologists and climatologists. For example, during years when the weather is not suitable for abundant grain production the price of wheat may be quite high.

16.3.2 Prewhitening

When checking for causality, the time series under investigation must first be prewhitened. Table 16.3.2 describes the types of models which were used to prewhiten the series from Table 16.3.1. In all cases, the models were determined by following the three stages of model construction in conjunction with the AIC (see Section 6.3) and in some instances the most appropriate models are constrained models for which some of the model parameters are omitted. For example, as explained in Sections 3.4.4 and 5.4.3, the best ARMA model for the sunspot series is a constrained ARMA (9,0) model where ϕ_3 to ϕ_8 are left out of the model and the original data are transformed by a square root transformation for which $\lambda = 0.5$ in [3.4.30] where x_t replaces z_t, and $c = 1$ due to some zero values in the series. Using the format in [16.2.3] or [3.4.4], the estimated sunspot model is written in difference equation form in [6.4.4] as

$$(1 - 1.245B + 0.524B^2 - 0.192B^9)(x_t - 10.673) = u_t \qquad [16.3.1]$$

where

$$x_t = (1/0.5)[(X_t + 1.0)^{0.5} - 1.0]$$

Notice for the Beveridge wheat price indices that the data are transformed using a natural logarithmic transformation where $\lambda = 0$ and $c = 0$ in [3.4.30]. The transformed data are then differenced once to remove nonstationarity by using [4.3.3] which is written as

Table 16.3.1. Time series used in the causality studies.

DATA SET	LOCATION	PERIOD	LENGTH
Sunspots	Sun	1700-1960	261
Annual Temperatures	English Midlands	1723-1970	248
12 Monthly Temperature Sequences	English Midlands	1723-1970	248 per month
St. Lawrence River	Ogdensburg, New York, USA	1860-1957	97
Volga River	Gorkii, USSR	1877-1935	58
Neumunas River	Smalininkai, USSR	1811-1943	132
Rhine River	Basle, Switzerland	1807-1957	150
Gota River	Sjotorp-Vanersburg, Sweden	1807-1957	150
Danube River	Orshava, Romania	1837-1957	120
Mississippi River	St. Louis, Missouri, USA	1861-1957	96
Beveridge Wheat Price Index	England	1500-1869	370
Tree Ring Widths	Campito Mountain, California, USA	1500-1969	470

$$y_t = \ln Y_{t+1} - \ln Y_t$$

for $t = 1,2,3, \ldots, n-1$. Following this, identification results explained in detail in Section 4.3.3 reveal that an ARMA (8,1) without ϕ_3 to ϕ_7 should be fitted to y_t where the estimated model is written using the notation of [16.2.4] as

$$(1 - 0.729B + 0.364B^2 + 0.119B^8)y_t = (1 - 0.783B)v_t \qquad [16.3.2]$$

The reader should bear in mind that only the family of ARMA models are entertained when selecting the best model to describe each data set in Table 16.3.2. In certain instances, it may be appropriate to also consider other types of models. For example, Akaike (1978) noted that because of the nature of sunspot activity a model based on some physical consideration of the generating mechanism may produce a better fit to the sunspot series than an ARMA model. For

Table 16.3.2. Models used to get residuals for the CCF studies.

DATA SET	ARMA (p,q) MODEL	$\hat{u}_t$	$\hat{v}_t$
Sunspots	(9,0) without ϕ_3 to ϕ_8, $\lambda = 0.5$ and c = 1	*	
Annual Temperature	(2,0) without ϕ_1	*	*
12 Monthly Temperature Sequences	(0,0) for all months	*	*
St. Lawrence River	(3,0) without ϕ_2 as in [6.4.2]		*
Volga River	(0,0)		*
Neumunas River	(0,1), $\lambda = 0$		*
Rhine River	(0,0)		*
Gota River	(2,0)		*
Danube River	(0,0)		*
Mississippi River	(0,1)		*
Beveridge Wheat Price Indices	(8,1) without ϕ_3 to ϕ_7, $\lambda = 0$, and series is differenced once		*
Tree Ring Widths	(4,0) without ϕ_2		*

modelling the sunspot series, Granger and Andersen (1978) utilized a bilinear model. Whatever the case, for each time series in Table 16.3.2 extensive diagnostic checking was executed to ensure that the best ARMA model was ultimately chosen.

16.3.3 Causality Studies

Following prewhitening, [16.2.6] is employed to calculate the residual CCF for two specified residual series. In the third and fourth columns of Table 16.3.2, *'s indicate when the residuals of a given series are used as $\hat{u}_t$ and/or $\hat{v}_t$, respectively, in [16.2.6]. Whenever two series are cross-correlated, the residual values are used for the time period during which the $\hat{u}_t$ and $\hat{v}_t$ data sets overlap. The sunspot residuals could possibly affect all the other series in Table 16.3.2 and, therefore, the sunspot residuals are separately cross-correlated with each of the remaining series in Table 16.3.2. For the monthly temperature data, each monthly sequence is considered as a separate sample when the residual CCF is calculated between the sunspot series as $\hat{u}_t$ and a given monthly temperature data set as $\hat{v}_t$. However, it is also possible that temperature can affect the phenomena listed below the temperature series in Table 16.3.2. For example, April temperatures may influence tree ring growth in the Northern Hemisphere since the month of April is when the growing season begins after the winter months. Consequently, the residual series for the annual temperature data set and the 12 monthly temperature sequences are each cross-correlated with the residual series of each of the data sets given below the temperatures.

In many situations, it may not be known whether or not one phenomenon definitely causes another. Although the direction of suspected causality is often known a priori due to a physical understanding of the problem, proper statistical methods must be employed to ascertain if the available evidence confirms or denies the presence of a significant causal relationship. Consider, for example, determining whether or not sunspots and riverflows are causally related. Obviously, it is only physically possible for sunspots to cause riverflows and not vice versa. Based upon ad hoc graphical procedures comparing annual flows of the Volga River in the USSR with yearly sunspot numbers, Smirnov (1969) postulated that sunspots unequivocally affect riverflows. However, when the residual CCF is used to detect scientifically causality, the results do not support Smirnov's strong claim. In Figure 16.3.1, the residual CCF along with the 95% confidence limits are presented for the residuals from the ARMA model fitted to the annual flows of the Volga River at Gorkii, USSR, and the residuals from the ARMA model fitted to the annual sunspot numbers (refer to Table 16.3.1 for a description of these data sets and to Table 16.3.2 for the types of models fitted to the two time series). As can be seen, there are no significant values of the residual CCF at lag zero and the smaller positive lags. If sunspot activity did affect the Volga flows, it would be expected that this would happen well within the time span of a few years. Therefore, the absence of significant values of the CCF from lags 0 to 2 or 3 indicates that the current information does not support the hypothesis that sunspots cause the Volga riverflows. The slightly large magnitudes at lags 5 and 11 are probably due to chance. Nevertheless, it is possible, but highly unlikely, that the value at lag 11 could be due to the fact that the best ARMA model could not completely remove the periodicity present in the sunspot series. Previously, Granger (1957) found that the periodicity of sunspot data follows a uniform distribution with a mean of about 11 years. However, the constrained sunspot model in [16.3.1] is designed to account for this. Moreover, the fitted model is subjected to rigorous diagnostic checks to demonstrate that the periodicity is not present in the model residuals and none of the values of the residual ACF are significantly different from zero, even at lag 11.

Besides the annual flows of the Volga River, no meaningful causality relationships are detected when the sunspot residuals are separately cross-correlated with the other riverflow residuals and also the remaining residuals series which are considered as $\hat{v}_t$ in Table 16.3.2. As emphasized earlier, if co.rect statistical procedures are not followed it would not be possible to reach the aforesaid conclusions regarding the causality relationships between the sunspots and the other phenomena. For example, in Figure 16.3.2 it can be seen that the values of the CCF calculated for the given annual sunspot and Gota riverflows series are large in magnitude at negative and positive lags (recall that the 95% confidence limits in Figure 16.3.2 are derived for independent series). Furthermore, the cyclic nature of the sunspot data is portrayed by the sinusoidal characteristics in the graph. To uncover the underlying causal relationship between the series it is necessary to examine the residual CCF. As just noted, the residual CCF for the

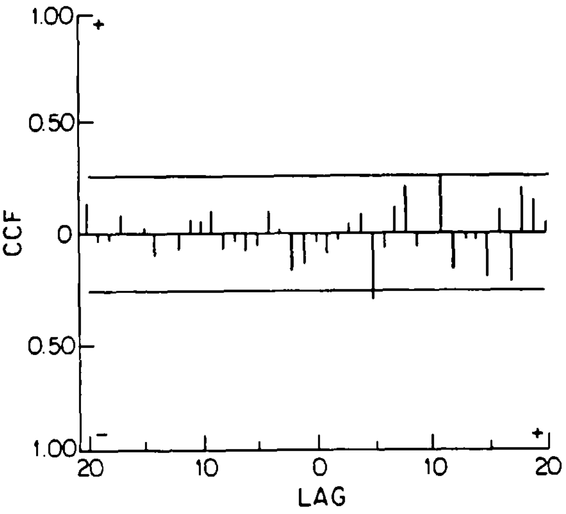


Figure 16.3.1. Residual CCF for the sunspot numbers and the Volga riverflows.

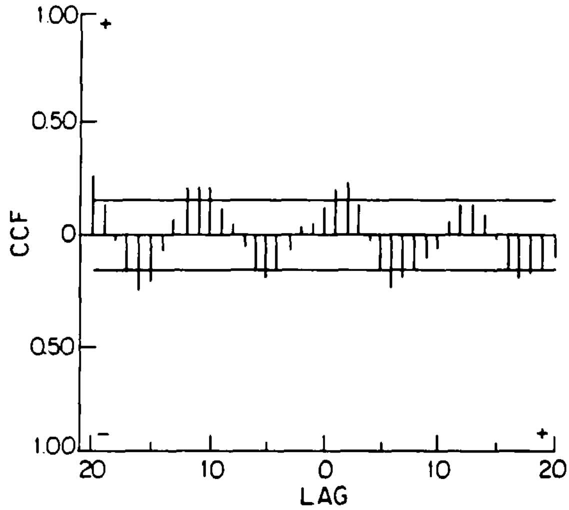


Figure 16.3.2. CCF for the given sunspot numbers and the Gota riverflows.

sunspot and Gota River series does not reveal that sunspot numbers affect the flows of the Gota River.

For the case where the $\hat{u}_t$ sequence, as represented by the residuals of the annual temperature data, is cross-correlated with each of the last nine $\hat{v}_t$ series in Table 16.3.2, no meaningful relationships are found. Nonetheless, some significant values of the residual CCF are discovered when each monthly temperature series is cross-correlated separately with each residual sequence for the riverflows and also the Beveridge wheat price indices. Table 16.3.3 shows the lags at which the residual CCF possesses large values when $\hat{u}_t$ is a designated monthly temperature series and $\hat{v}_t$ is either the annual riverflow or Beveridge wheat price index residuals. Since it would be expected from a physical viewpoint that a given monthly temperature data set would have the most effect upon the other time series in the current year or perhaps one or two years into the future, large values of the residual CCF are only indicated in Table 16.3.3 when they occur at lags 0 to 2. As an illustrative example, consider the graph of the residual CCF for the August temperatures and the Gota River residuals which is shown in Figure 16.3.3. As can be seen, the large negative correlation at lag zero extends well beyond the 95% confidence limits. When the Q_L statistic in [16.2.7] is calculated for lags 0 to 2, the estimated value for the residual CCF in Figure 16.3.3 is 26.6. Because this value is much larger than the tabulated $\chi^2(3)$ value of 7.8 for the 5% significance level, one must reject the null hypothesis that the August temperatures do not affect the annual flows of the Gota River.

It would be expected that temperature could significantly affect tree ring growth. As noted by La Marche (1974), because Bristlecone Pines are located at the upper treeline on mountains, temperature is a key factor in controlling growth. However, this growth would only be sensitive to local temperature conditions and the temperatures recorded in the English Midlands are probably not representative of the temperatures at Campito Mountain in California. If local temperatures were available, the residual CCF between the local temperatures and tree ring widths could be calculated to ascertain the type of causality which is present.

16.4 CONCLUSIONS

Comprehensive procedures are now available for detecting causal relationships between two time series. The results of Table 16.3.3 demonstrate that monthly temperatures can significantly affect annual riverflows and also the price of wheat. However, no meaningful links are found between the annual sunspot numbers and the other phenomena designated by $\hat{v}_t$ in Table 16.3.2. In particular, the statistical evidence from Figure 16.3.1 cannot support the claim (Smirnov, 1969) that sunspots significantly affect the annual flows of the Volga River. While some of the findings of Section 16.3 may be somewhat interesting, it is also informative to note the types of results that Pierce (1977) discovered in the field of econometrics. Using residual CCF studies, Pierce found that numerous economic variables which were generally regarded by economists as being strongly interrelated were in fact independent or else only weakly correlated. These conclusions are of course based upon the information included in the time series which Pierce analyzed. If it were possible to improve the design of the data collection scheme for a causality study, this would of course enhance the conclusions reached at the analysis stage. Certainly, it is necessary that a sufficiently wide range of values of the relevant variables appear in the sample in order to increase the probability of detecting relationships which do actually exist in the real world. However, as is the case in economics and also in the natural sciences, the experimenter

Table 16.3.3. Residual CCF results of monthly temperature and other series.

$\hat{v}_t$	MONTHLY TEMPERATURES $\hat{u}_t$	LAGS FOR LARGER VALUES OF RESIDUAL CCF
St. Lawrence River	February	1
Volga River	February	2
	April	2
	May	0 and 1
	July	2
Neumunas River	May	0
	July	2
	December	2
Rhine River	October	0
Gota River	June	0
	July	0
	August	0
	September	0
Danube River	September	0
	October	0
Mississippi River	December	1
Beveridge Wheat Price Index	February	0
	November	0
	December	0

has little control over the phenomena which produce the observations and must therefore be content with the data that can be realistically collected. Perhaps *God* may have a switch that can greatly vary the number of sunspots that appear on the sun so that *mortal man* can assess beyond a shadow of a doubt whether or not sunspots can significantly affect riverflows.

Given the available information, it is essential that the data be properly analyzed. For example, if a sample CCF were calculated for the x_t and y_t series, spurious correlations may seem to indicate that the variables are causally related (see Figure 16.3.2, for instance).

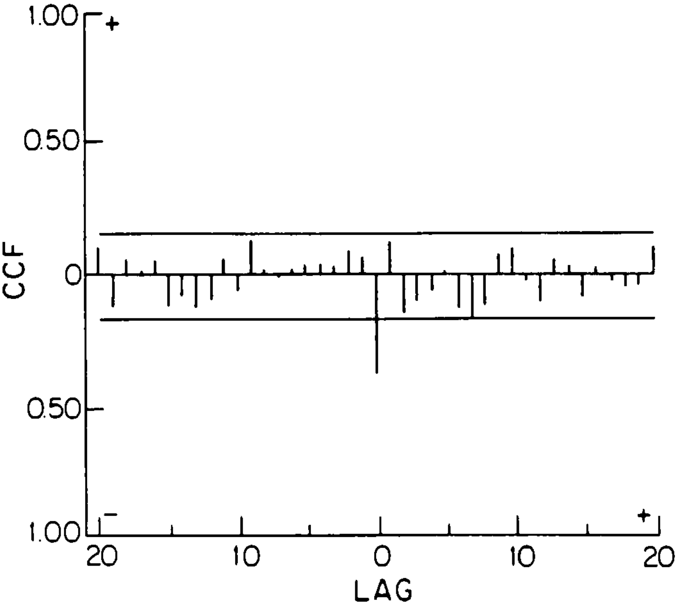


Figure 16.3.3. Residual CCF for the August temperatures and the Gota riverflows.

However, an examination of the residual CCF for the two series may clearly reveal that based upon the given data no meaningful relationships do in fact exist between the two phenomena. It is of course possible that no significant correlations may appear in the residual CCF even though two variables are functionally related. This is because correlation is only a measure of linear association and nonlinear relationships that contain no linear component, may be missed. To minimize the occurrence of this type of error, the fitted ARMA models that are used to prewhiten the series are subjected to stringent diagnostic checks. In this way, any problems that arise due to the use of these linear models will be detected prior to examining the residual CCF.

Subsequent to the revelation of causality using the residual CCF, a dynamic model can be built to describe mathematically the formal connections between the x_t and y_t series. In most hydrological and other geophysical applications, usually one variable causes another and there is no feedback. For instance, precipitation causes riverflows and this unidirectional causality cannot be reversed. In terms of the residual CCF, for unidirectional causality from X to Y, the residual CCF is nonzero at one or more lags for $k > 0$, $\rho_{uv}(0)$ may be either zero or have some nonzero value, and the value of the residual CCF at all negative lags must be zero (see Table 16.2.1). To describe mathematically the formal connections between the x_t and y_t series, the TFN model described in the next chapter constitutes a flexible dynamic model which can be utilized. An inherent advantage of TFN models is that well developed methodologies are available for use at the three stages of model construction. For instance, at the identification step the results of the residual CCF study that detected the causal relationship in the first place, can be utilized to design the dynamic model (Haugh and Box, 1977). When the y_t series has been altered by one or more external interventions, then intervention components can be introduced into the TFN model to account for possible changes in the mean level (see Chapters 19 and 22).

When there is feedback between X and Y, Table 16.2.1 shows that $\rho_{uv}(k)$ is nonzero at both positive and negative lags. The multivariate models in Chapters 20 and 21 are the type of dynamic models which can be used to model rigorously the dynamical characteristics of the feedback. Nevertheless, the reader should keep in mind that TFN models are used much more than multivariate models in hydrology and environmental engineering, since most natural systems do not possess feedback. Consequently, TFN models are described in more detail than multivariate models within this text.

PROBLEMS

16.1 Granger causality is defined in Section 16.2.1. Explain at least one other way in which scientists define causality between two phenomena. You may, for instance, wish to examine the path analysis procedure for studying relationships among variables which Kaplan and Thode (1981) apply to water resources data. Another procedure to consider for investigating causality is the coherence function (Bendat and Piersol, 1980) mentioned at the end of Section 6.2.2. Compare the residual CCF method to the other techniques for causality detection in terms of similarities and differences in the basic procedures, as well as advantages and drawbacks.

16.2 As is illustrated in Figure 16.3.2, spurious relationships between two variables can be found by improperly comparing the two variables. One way to overcoming spurious statistical connections between two time series is to employ the residual CCF approach of section 16.2.2. Find a statistical study in a field which is of interest to you where you think that scientists may have discovered spurious causal connections between two variables which do not really exist. Point out where the authors followed an improper procedure and explain how it can be corrected.

16.3 Select two annual time series for which you suspect one variable causes the other. For instance, you may have a representative yearly regional precipitation series which causes average annual riverflows in a river falling within the region. For these two data sets, carry out the following tasks:

(a) Prewhiten each series by fitting an ARMA to the series and thereby obtaining the model residuals.

(b) Calculate and plot the residual CCF for the two series along with the 95% confidence limits.

(c) Describe the stochastic relationships that you find in part (b). Explain why your findings make sense by linking them with the physical characteristics of the system under study. If, for example, you are examining a hydrological system, include aspects of the hydrological cycle described in Section 1.5.2 in your explanation.

16.4 Design and calibrate a TFN model for formally describing the dynamic relationships between the two time series examined in problem 16.3. Perform diagnostic checks to ensure that your fitted model adequately links the two data sets.

16.5 Choose a set of 6 to 10 time series in a field which you are working. Following the approach employed for the time series in Section 16.3, carry out a systematic causality study among your data sets. Comment upon the interesting results that you discover.

16.6 Select two seasonal time series, such as average monthly precipitation and riverflows, for which it makes sense to remove the seasonality by employing a suitable deseasonalization technique from Section 13.2.2. After fitting an ARMA model to each of the deseasonalized series, carry out a causality study to examine the relationships among these series.

REFERENCES

DATA SETS

Beveridge, W. H. (1921). Weather and harvest cycles. *Economics Journal*, 31:429-552.

La Marche, Jr., V. C. (1974). Paleoclimactic inferences from long tree-ring records. *Science*, 183:1042-1048.

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society*, 79:242-261.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

CAUSALITY

Bartlett, M. S. (1935). *Stochastic Processes*. Cambridge University Press, London.

Bendat, J. S. and Piersol, A. G. (1980). *Engineering Applications of Correlation and Spectral Analysis*. Wiley, New York.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26:211-252.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424-438.

Haugh, L. D. (1972). *The Identification of Time Series Interrelationships with Special Reference to Dynamic Regression*. Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison, Wisconsin.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378-385.

Hipel, K. W., McLeod, A. I. and Li, W. K. (1985). Causal and dynamic relationships between natural phenomena. In Anderson, O. D., Ord, J. K. and Robinson, E. A., Editors, *Time Series Analysis: Theory and Practice*, pages 13-34. North-Holland, Amsterdam.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Kaplan, E. and Thode Jr., H. C. (1981). Water quality, energy and socioeconomics: path analyses for studies of causality. *Water Resources Research*, 17(3):491-503.

McLeod, A. I. (1979). Distribution of the residual cross-correlation in univariate ARMA time series models. *Journal of the American Statistical Association*, 74(368):849-855.

Pierce, D. A. (1977). Relationships - and the lack thereof - between economic time series, with special reference to money and interest rates. *Journal of the American Statistical Association*, 72(357):11-21.

Pierce, D. A. and Haugh, L. D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5:265-293.

Pierce, D. A. and Haugh, L. D. (1979). The characterization of instantaneous causality, A comment. *Journal of Econometrics*, 10:257-259.

Price, J. M. (1979). The characterization of instantaneous causality, A correction. *Journal of Econometrics*, 10:253-256.

Stedinger, J. R. (1981). Estimating correlations in multivariate streamflow models. *Water Resources Research*, 17(1):200-208.

Wiener, N. (1956). The theory of prediction. In Beckenback, E., Editor, *Modern Mathematics for Engineers*, Series 1, Chapter 8, McGraw-Hill, New York.

SUNSPOT NUMBERS

Akaike, H. (1978). On the likelihood of a time series model. Paper presented at the Institute of Statisticians 1978 Conference on Time Series Analysis and Forecasting, Cambridge University.

Brillinger, D. R. (1969). A search for a relationship between monthly sunspot numbers and certain climatic series. *Bulletin of the International Statistical Institute*, 43:293-307.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysics Journal*, 126:152-158.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Gottingen.

Rodriquez-Iturbe, I. and Yevjevich, V. M. (1968). The investigation of relationship between hydrologic time series and sunspot numbers, hydrology paper no. 26. Technical report, Colorado State University, Fort Collins, Colorado.

Smirnov, N. P. (1969). Causes of long-period streamflow fluctuations. *Bulletin of the All-Union Geographic Society (Izvestiya VGO)*, 101(5):443-440.

TRANSFER FUNCTION NOISE MODELLING

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Haugh, L. D. and Box, G. E. P. (1977). Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association*, 72(357):121-130.

CHAPTER 17

CONSTRUCTING TRANSFER FUNCTION-NOISE MODELS

17.1 INTRODUCTION

Hydrologists and other types of natural scientists often require a stochastic model which realistically describes the dynamic relationships connecting a single output series with one or more input series. For example, a stochastic model can be developed for formally modelling the mathematical linkage of a single output or response series such as seasonal riverflows to one or more input or covariate series such as precipitation and temperature. When inputs are incorporated into a stochastic model, the manner in which the input dynamically affects the output over time is mathematically modelled. For this reason, stochastic models which possess inputs are often referred to as *dynamic models*. Because the dynamic characteristics of the physical system being modelled are incorporated into the overall stochastic or dynamic model, this allows for more accurate forecasts to be made (see Chapter 18) and for more realistic values to be generated in simulation studies. Consequently, the main purpose of this chapter is to describe a flexible dynamic model which can be applied to many kinds of environmental problems where the data can be either nonseasonal or seasonal. The particular kind of dynamic model which is studied is called a *transfer function-noise model* where the acronym TFN is used for denoting transfer function-noise.

Throughout the chapter, *practical applications* are employed for clearly explaining how TFN models can be easily constructed by following the identification, estimation and diagnostic check stages of *model construction*. In particular, for model identification it is clearly pointed out how both a sound physical understanding of the problem being studied and comprehensive statistical procedures can be employed for designing an appropriate dynamic model to fit to a given set of time series. When designing a suitable dynamic model, usually it is most instructive and convenient to consider simpler models and to gradually increase the complexity of the model until a reasonable model is built. Accordingly, TFN models with a single input are entertained in the next section. The ways in which the residual CCF from Section 16.2.2, as well as two other techniques, can be used in model identification are thoroughly explained in Section 17.3 along with other model building techniques. For the application in Section 17.4.2 for a dynamic model having a single input, a TFN model is developed for relating average monthly upstream riverflows to downstream flows. Also, because the causality studies in Chapter 16 indicate that the average August temperature significantly affect the average annual flows of the Gota River, a dynamic model is constructed in Section 17.4.3 for rigorously describing this relationship.

For the situation where there is more than one input series, the efficacy of the model building techniques is clearly demonstrated in Section 17.5.4 by designing a TFN model to describe the dynamic relationships connecting a monthly river flow series in Canada to precipitation and temperature covariate series. In the process of selecting the most appropriate model to fit to the series, a number of useful modelling procedures are suggested. Often there are more than one precipitation and temperature series and a statistical procedure is presented for creating a single sequence to represent the precipitation or temperature series. This approach can be utilized in place of the rather ad hoc methods such as the Isohyetal and Thiessen polygon techniques (see Viessman et al. (1977) for a description of the Isohyetal and Thiessen polygon methods). To decide upon which series to include in the dynamic model and also design the form of the transfer function connecting a covariate series to the output, cross-correlation analyses of the residuals of models fitted to the series can be employed. The TFN model which is ultimately chosen by following contemporary modelling procedures can be used for applications such as forecasting and simulation and also providing insights into the physical characteristics of the phenomena being examined.

A TFN model is ideally designed for reflecting the physical characteristics of many natural systems. In a watershed, for example, the two basic equations that govern flow are the continuity equation or conservation of mass and the conservation of momentum. Delleur (1986) demonstrates that a TFN model can represent the discrete-time version of the continuous-time differential equation that is derived from the above two conservation principles. Because a TFN model can be readily calibrated by following the three stages of model construction given in this chapter and also the model reflects the physical characteristics of the system being described, the TFN model is ideally suited for real-world applications in hydrology and other sciences. For a discussion of the *physical justification* of ARMA models, the reader may wish to refer to Section 3.6.

Besides the TFN modelling applications presented in Sections 17.4.2, 17.4.3 and 17.5.4, other case studies are presented in Chapter 18 with the forecasting experiments. The intervention model, which in reality constitutes a special class of TFN models for modelling the affects of external interventions upon the mean level of the output series, can also handle multiple covariate series. Applications of intervention modelling for which there are input series are presented in Sections 19.5.4 and 22.4.2. Because TFN models work so well in practice, there are many published case studies of TFN in water resources, environmental engineering as well as other fields. Some of the many TFN modelling applications in the physical sciences include contributions in hydrology (Anselmo and Ubertini, 1979; Baracos et al., 1981; Hipel et al., 1992; Chow et al., 1983; Thompstone et al., 1983; Snorrason et al., 1984; Hipel et al., 1985; Maidment et al., 1985; Olason and Watt, 1986; Fay et al., 1987; Gurnell and Fenn, 1984; Lemke, 1990, 1991), dendroclimatology (Li, 1981, Ch. 8), modelling wastewater treatment plants (Capodaglio et al., 1992), and fish population studies (Noakes et al., 1987; Campbell et al., 1991; Welch and Noakes, 1991). Finally, a stochastic model closely related to the TFN model is defined and evaluated in Section 17.6.

17.2 TRANSFER FUNCTION-NOISE MODELS WITH A SINGLE INPUT

17.2.1 Introduction

For many natural systems it is known a priori if one variable, or set of variables, causes another. Lake levels, for instance, are obviously affected by precipitation. In situations where it is uncertain if one physical phenomenon causes another, the residual CCF discussed in Section 16.2.2 can be employed. For example, one may wish to find out if sunspots cause riverflows, as is done in Section 16.3.3. Whatever the case, subsequent to establishing the existence and direction of causality between two phenomena, a TFN model can be built to model mathematically the dynamical characteristics of the physical system.

Consider the situation where a variable X causes a variable Y. Let the set of observations for the variables X and Y be represented by the series X_t and Y_t , respectively, for $t=0,1,2,\ldots,n$. The X_t and Y_t series may be transformed using a transformation such as the Box-Cox transformation in [3.4.30] to form the x_t and y_t series, respectively. When the data are seasonal, the seasonality in each series may be removed using the deseasonalization techniques presented in Section 13.2.2. No matter what kinds of transformations are performed before fitting the dynamic model, x_t and y_t will always be used to represent the transformed series.

Qualitatively, a TFN model can be written as

output = dynamic component + noise

The manner in which the input, x_t , dynamically affects the output, y_t, is modelled by the dynamic component. However, usually influences other than the input variable X will also affect Y. The accumulative effect of other such influences is called the noise or disturbance where the noise is usually correlated and may be modelled by an ARMA or ARIMA model. The dynamic and noise components are now discussed separately.

17.2.2 Dynamic Component

The dynamic relationship between X and Y can be modelled by a transfer function model as

$$\begin{aligned} y_t &= v_0x_t + v_1x_{t-1} + v_2x_{t-2} + \cdots \\ &= v(B)x_t \end{aligned}$$

where $v(B) = v_0 + v_1B + v_2B^2 + \cdots$, is referred to as the transfer function and the coefficients, $v_0, v_1, v_2, \ldots,$ are called the *impulse response function* or *impulse response weights*. When there are nonzero means μ_y and μ_x for the y_t and x_t series, respectively, the transfer function model can be written in terms of deviations from the mean level as

$$y_t - \mu_y = v(B)(x_t - \mu_x)$$

The deterministic transfer function models how present and past values of X affect the current value of Y. The x_t series is often referred to as the *input*, *covariate* or *exogenous series*, while the y_t series is called the *output*, *response* or *endogenous series*.

Recall from [3.4.18], that for an ARMA model, the infinite MA operator is written in terms of two finite operators as

$$\Psi(B) = \frac{\theta(B)}{\phi(B)}$$

where $\Psi(B)$ is the infinite MA operator, $\theta(B)$ is the MA operator of order q and $\phi(B)$ is the AR operators of order p. In practical applications, only a small number of AR and MA parameters are required to model a given series. Hence, the infinite MA operator $\Psi(B)$ can be parsimoniously represented by $\frac{\theta(B)}{\phi(B)}$. In a similar fashion, the *transfer function*, $v(B)$, can be economically written as

$$v(B) = \frac{\omega(B)}{\delta(B)} = \frac{\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_m B^m}{1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r}$$

where $\omega(B)$ is the operator of order m in the numerator of the transfer function and $\omega_0, \omega_1, \cdots, \omega_m$, are the parameters of $\omega(B)$; and $\delta(B)$ is the operator of order r in the denominator of the transfer function and $\delta_1, \delta_2, \ldots, \delta_r$, are the parameters of $\delta(B)$.

If a system is *stable*, a finite incremental change in the input results in a finite incremental change in the output. In other words, a bounded change in the covariate variable causes a bounded change in the response. For the case of the transfer function model, this implies that the infinite series $v_0 + v_1 B + v_2 B^2 + \ldots$, converges for $|B| \leq 1$. Because the convergence of $v(B)$ is controlled by the operator $\delta(B)$ which is in the denominator of $v(B)$, the *requirement of stability* is that the roots of the characteristic equation $\delta(B) = 0$ lie outside the unit circle (Box and Jenkins, 1976, Ch. 10). Notice that the requirement of stability for a discrete transfer function model is analogous to the stationarity requirement for an ARMA model in Section 3.4.2 where the roots of $\phi(B) = 0$ must lie outside the unit circle.

In some situations, there may be a *delay time* before X affects Y. For instance, when an organic pollutant is discharged into a river, there may be a delay time before certain biological processes take place and the dissolved oxygen level of the river drops. If this delay time is denoted by b, where b is a positive integer for use in a model using evenly spaced discrete time points, the transfer function model can be written as

$$y_t - \mu_y = v(B)(x_t - \mu_x) = \frac{\omega(B)}{\delta(B)} B^b (x_t - \mu_x) \qquad [17.2.1]$$

When the parameters for the $\omega(B)$ and $\delta(B)$ operators are known, as is the case when they are estimated from the given data, the v_k coefficients can be determined by equating coefficients of B^k in

$$\delta(B)v(B) = \omega(B)B^b$$

The v_k weights can be more conveniently determined by expressing the above equation as

$$\delta(B)v_{k+b} = -\omega_k \quad \text{for } k = 1,2,\ldots \qquad [17.2.2]$$

where B operates on the subscript in the v_{k+b} coefficient and therefore $B^j v_{k+b} = v_{k+b-j}$; $v_b = \omega_0$ and $v_k = 0$ for $k < b$; $\omega_k = 0$ for $k < 0$ and $k > m$.

As an example of how to use [17.2.2], consider the situation where $v(B) = \frac{\omega_0 B}{1-\delta_1 B}$ and it is necessary to determine the impulse response function where ω_0 and δ_1 are given. For this transfer function, [17.2.2] becomes

$$(1 - \delta_1 B)v_{k+1} = -\omega_k$$

Because there is a delay factor of one, $v_0 = 0$ and $v_1 = \omega_0$. When k=1 in the above equation

$$(1 - \delta_1 B)\nu_2 = -\omega_1 = 0$$

or

$$\nu_2 - \delta_1 \nu_1 = 0$$

Therefore,

$$\nu_2 = \delta_1 \nu_1 = \delta_1 \omega_0$$

For k=2

$$(1 - \delta_1 B)\nu_3 = -\omega_2 = 0$$

$$\nu_3 - \delta_1 \nu_2 = 0$$

$$\nu_3 = \delta_1 \nu_2 = \delta_1^2 \omega_0$$

In general,

$$\nu_k = \delta_1^{k-1} \omega_0$$

Because the ν_k weights are calculated in a similar fashion to the ψ_k coefficients in the infinite MA operator for an ARMA model, the reader can refer to Section 3.4.3 for further examples of how to determine the impulse response weights.

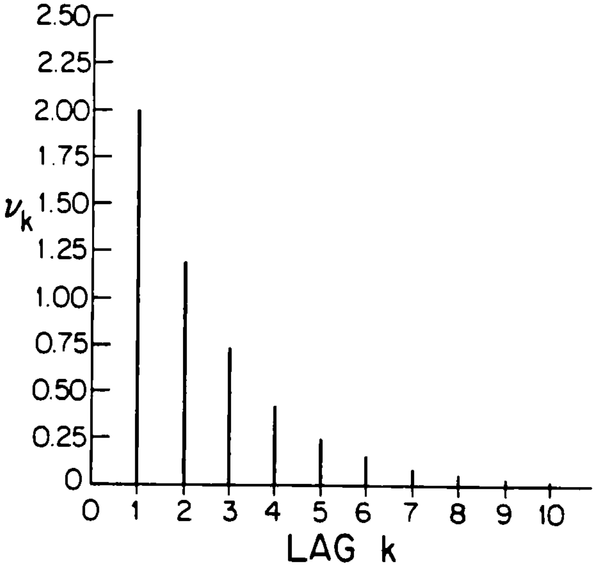


Figure 17.2.1. Impulse response function for $\omega_0 = 2$ and $\delta_1 = 0.6$.

When the calculated impulse response function is plotted for the case where $\omega_0 = 2$ and $\delta_1 = 0.6$, the impulse response function is as shown in Figure 17.2.1. Notice that $v_0 = 0$ due to the delay factor of one in the transfer function $v(B) = \frac{\omega_0 B}{(1-\delta_1 B)}$ and that the impulse response function attenuates from lag 2 onwards because of the operator $(1-\delta_1 B)$ in the denominator of the transfer function. The general form of the transfer function, especially the operator $\delta(B)$ in the denominator of the transfer function, allows for great flexibility in the design of a transfer function for ascertaining the effects of the input upon the output. For example, when a variable X affects Y after a delay factor of one and the magnitude of the effect upon y_t of each x_t decreases more and more for $t = t+1, t+2, t+3, \ldots,$ then a transfer function of the form $v(B) = \frac{\omega_0 B}{(1-\delta_1 B)}$ may be appropriate.

If the input, x_t, in [17.2.1] is indefinitely held at some fixed value for a stable system, the output, y_t, will eventually reach an equilibrium point which is called the steady state level. Using the form of the transfer function model in [17.2.1], the steady state relationship is

$$y_t - \mu_y = g(x_t - \mu_x)$$

where g is called the steady state gain when x_t is held indefinitely at a fixed level. Suppose that $(x_t - \mu_x)$, which is the deviation of the input from the mean level, is kept at a value of +1. Then

$$\begin{aligned} y_t - \mu_y &= g1 \\ &= (v_0 + v_1 B + v_2 B^2 + \ldots)1 \\ &= v_0 + v_1 + v_2 + \cdots \end{aligned}$$

Consequently, for a stable system, the *steady state gain* is

$$g = \sum_{i=0}^{\infty} v_i = \frac{\omega(1)}{\delta(1)} \qquad [17.2.3]$$

For the case where

$$v(B) = \frac{\omega_0 B}{(1-\delta_1 B)} = \omega_0(1 + \delta_1 B + \delta_1^2 B^2 + \cdots)B$$

in [17.2.1], the steady state gain is simply

$$g = \frac{\omega_0}{1 - \delta_1}$$

by substituting $B = 1$ into the equation.

17.2.3 Noise Term

In practice, a physical system cannot be realistically modelled by using only the deterministic transfer function model in [17.2.1]. As shown qualitatively below, usually there is noise left in the system after the deterministic dynamic effects of the input upon the output have been accounted for.

output – dynamic component = noise

Because the noise term, which is denoted by N_t , is often autocorrelated and not white, it can be conveniently modelled using the ARMA model in [3.4.4], as

$$\phi(B)N_t = \theta(B)a_t \qquad [17.2.4]$$

or

$$N_t = \frac{\theta(B)}{\phi(B)}a_t$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA operators of orders p and q, respectively, and a_t is the white noise which is $IID(0,\sigma_a^2)$. Although the a_t series can be assumed to be $\mathrm{IID}(0,\sigma_a^2)$ from a theoretical viewpoint, for practical reasons the a_t must be assumed to follow a given distribution in order to be able to obtain estimates for the model parameters. Consequently, the a_t are assumed to be $NID(0,\sigma_a^2)$ and, as demonstrated by practical applications, this assumption does not restrict the flexibility of the TFN model in any way. In fact, if the a_t series were not assumed to be normally distributed, it would be virtually impossible to obtain efficient estimates for the model parameters. As shown by the examples in this chapter and also Chapters 18, 19 and 22, for most environmental applications the noise term is stationary and therefore can be modelled using an ARMA model. Nevertheless, if the noise were nonstationary and differencing were required, it could be modelled using an ARIMA model.

17.2.4 Transfer Function-Noise Model

By combining [17.2.4] and [17.2.1], an overall TFN model for simultaneously modelling both the dynamic characteristics and noise contained in the system is formed as

$$\begin{aligned} y_t - \mu_y &= v(B)(x_t - \mu_x) + N_t \\ &= \frac{\omega(B)}{\delta(B)}B^b(x_t - \mu_x) + \frac{\theta(B)}{\phi(B)}a_t \qquad [17.2.5] \end{aligned}$$

Because the above model possesses both dynamic and stochastic or noise components, it could perhaps be referred to as a stochastic dynamic model. However, since all of the models discussed in this text are stochastic in nature, usually this TFN model is simply referred to as a dynamic model. Also, because the noise term models what the dynamic component cannot account for, it is assumed that N_t is independent of x_t in [17.2.5]. Since N_t is generated by a_t and x_t is generated by u_t in [16.2.3], this in turn means that the a_t series in [17.2.5] is independent of the u_t series in [16.2.3].

As noted by authors such as Abraham and Ledolter (1983, p. 338) and Vandaele (1983, pp. 263-264), the TFN model can be thought of as a generalization of the usual regression model for various reasons. Firstly, the model in [17.2.5] allows for the noise term to be correlated whereas the error term in a regression model is assumed to be white. Secondly, due to the flexible manner in which the transfer function in [17.2.5] is designed with an operator in both the denominator and numerator, the dynamic relationships between the output and input are more realistically modelled with a TFN model. As pointed out by Young (1984, p. 104), linear regression models are primarily utilized in the evaluation of static relationships among variables and are not generally suitable for use in dynamic systems analysis. Finally, as is also the case with regression analysis, one or more input series can be incorporated into a TFN model.

17.3 MODEL CONSTRUCTION FOR TRANSFER FUNCTION-NOISE MODELS WITH ONE INPUT

17.3.1 Model Identification

No matter what type of model is being fitted to a given data set it is recommended to follow the identification, estimation and diagnostic check stages of model development. As is the case for most of the models in this book, the three stages of model construction for a TFN model closely parallel those already described for models such as the ARMA model (see Chapters 5 to 7) and the seasonal models (Part VI). Nevertheless, some additional model building techniques are required, especially for model identification. When designing a TFN model, the number of parameters required in each of the operators contained in the dynamic and noise transfer functions in [17.2.5] must be identified. The three procedures for model identification described in this chapter are the empirical approach that has been used when modelling hydrological time series (Hipel et al., 1977, 1982, 1985), the technique of Haugh and Box (1977) which uses the residual CCF of Section 16.2.2, and the method of Box and Jenkins (1976) which is based upon suggestions by Bartlett (1935). The latter two methodologies rely heavily upon the results of cross-correlation studies and often the first procedure can be used in conjunction with either the second or third approaches. Subsequent to a description of the identification and other model building procedures, two practical applications are employed in Section 17.4 for demonstrating how they work in practice. These and other practical applications show that the empirical identification approach is usually the simplest to use in practice.

Empirical Identification Approach

The two major steps involved in model identification by the empirical method are as follows:

(i) Based upon an understanding of the physical phenomena that generated the time series and also the mathematical properties of the TFN model, identify the transfer function $v(B)$ in [17.2.5]. For example, when modelling a monthly time series where the input is a precipitation time series and the output is a riverflow sequence, it may be known from the physical characteristics of the watershed that the rainfall for the current month only causes direct runoff during the present time period and one month into the future. Therefore, it may be appropriate to employ in [17.2.5] a transfer function of the form

$$v(B) = \omega_0 - \omega_1 B = v_0 + v_1 B$$

When the present value of the covariate series causes an immediate change in the output, y_t, where this change attenuates as time progresses, a suitable transfer function may be

$$v(B) = \frac{\omega_0}{(1-\delta_1 B)} = \omega_0(1 + \delta_1 B + \delta_1^2 B^2 + \cdots)$$

$$= v_0 + v_1 B + v_1 B^2 + \cdots$$

The v_i coefficients decrease in value for increasing lag since $|\delta_1| < 1$ in order for the root of $(1 - \delta_1 B) = 0$ to lie outside the unit circle. Hipel et al. (1977, 1982, 1985) present a number of practical applications where a physical understanding of the process in conjunction with some simple graphical identification techniques are utilized to design transfer functions for covariate series and also intervention series which are needed when the effects of external interventions upon the response series must be incorporated into the model (see Chapters 19 and 22).

(ii) After deciding upon the form of $v(B)$, identify the parameters needed in the noise term in [17.2.5]. To accomplish this, first fit the model in [17.2.5] to the series where it is assumed that the noise term is white and consequently the TFN model has the form

$$y_t - \mu_y = v(B)(x_t - \mu_x) + a_t$$

In practice, the noise term is usually correlated. Therefore, after obtaining the estimated residual series, $\hat{a}_t$, for the above model, the type of ARMA model to fit to the calculated noise series can be ascertained by following the usual procedures of model development for a single series described in Chapters 5 to 7. By using the identified form of N_t for the noise term along with the previously designed dynamic component, the TFN model in [17.2.5] is now completely specified.

Haugh and Box Identification Method

It seems logical that a model similar to the one in [17.2.5] that relates the x_t and y_t series, could be developed for connecting their innovation series given as μ_t and v_t, respectively. Recall from Section 16.2.2 that the μ_t disturbances are for the ARMA model fitted to the x_t series as

$$\phi_x(B)(x_t - \mu_x) = \theta_x(B)u_t$$

in [16.2.3]. Likewise, the v_t innovations are formed by prewhitening the y_t series using the ARMA model written as

$$\phi_y(B)(y_t - \mu_y) = \theta_y(B)v_t$$

in [16.2.4]. As shown by Haugh and Box (1977), a TFN model for the residual series can be written as

$$v_t = \nu'(B)u_t + N'_t \qquad [17.3.1]$$

where $\nu'(B) = \dfrac{\omega'(B)}{\delta'(B)}$ is the transfer function which has the same form as the transfer function in [17.2.5] except that a prime symbol is assigned to each parameter; $N'_t = \dfrac{\theta'(B)}{\phi'(B)} a_t$ is the ARMA noise term where the MA operator, $\theta'(B)$, and the AR operator, $\phi'(B)$, are designed in the same fashion as their counterparts in [3.4.4], [16.2.3] or [16.2.4]. In their paper, Haugh and Box (1977) derive the relationships between the transfer functions in [17.2.5] and [17.3.1] as

$$\nu(B) = \frac{\omega(B)}{\delta(B)} = \frac{\theta_y(B)}{\phi_y(B)} \frac{\omega'(B)}{\delta'(B)} \frac{\phi_x(B)}{\theta_x(B)} \qquad [17.3.2]$$

and

$$\frac{\theta(B)}{\phi(B)} = \frac{\theta_y(B)}{\phi_y(B)} \frac{\theta'(B)}{\phi'(B)} \qquad [17.3.3]$$

By knowing the form of the operators on the right hand sides of [17.3.2] and [17.3.3], the parameters needed for the model in [17.2.5] can be identified. The detailed steps required for executing this identification process are as follows, where the first two steps follow the development of the residual CCF described in Section 16.2.2:

(i) Determine the most appropriate ARMA models to fit to the x_t and also the y_t series by adhering to the three stages of model construction. In addition to the ARMA model parameters, estimates are also obtained for the innovation series in [16.2.3] and [16.2.4] at the estimation stage of model development. The estimated innovation or residual series, $\hat{u}_t$ and $\hat{v}_t$, are formed from the process of prewhitening the given x_t and y_t series in [16.2.3] and [16.2.4], respectively using the calibrated ARMA models.

(ii) By utilizing [16.2.6], calculate the residual CCF for the $\hat{u}_t$ and $\hat{v}_t$ series, along with the 95% confidence limits.

(iii) Based upon the characteristics of the residual CCF identify the parameters needed in the transfer function $\nu'(B)$ in [17.3.1]. As demonstrated by Haugh and Box (1977, p. 126), the theoretical CCF for u_t and v_t is directly proportional to the impulse response function given by $\nu'_0, \nu'_1, \nu'_2, \cdots$, here

$$\nu'(B) = \frac{\omega'(B)}{\delta'(B)} = \nu'_0 + \nu'_1 B + \nu'_2 B^2 + \cdots$$

is the transfer function in [17.3.1]. In particular, the kth parameter in $\nu'(B)$ is related to the theoretical CCF by the equation

$$\nu'_k = \sigma_u \sigma_v \rho_{uv}(k), \quad k = 0,1,2,\cdots \qquad [17.3.4]$$

where σ_u and σ_v are the theoretical standard deviations for μ_t and ν_t, respectively, while $\rho_{uv}(k)$ is the theoretical residual CCF in [16.2.5]. Therefore, the form of the transfer function can be identified from the residual CCF and when the quantities on the right hand side of [17.3.4] are replaced by their estimates in [16.2.6], an initial estimate can be obtained for ν'_k. If, for example, the residual CCF possesses values which are significantly different from zero only at lags 0, 1, and 2, the transfer function may be identified as

$$v'(B) = \omega'_0 - \omega'_1 B - \omega'_2 B^2 = v'_0 + v'_1 B + v'_2 B^2$$

When the value of the residual CCF at lag 0 is significantly different from zero, the values attenuate for increasing positive lags, and the residual CCF values at negative lags are not significant, then the impulse response function must be designed to mimic this behaviour. Accordingly, an appropriate transfer function may be

$$v'(B) = \frac{\omega'_0}{(1-\delta'_1 B)} = \omega'_0(1 + \delta'_1 B + \delta'^2_1 B^2 + ...)$$

$$= v'_0 + v'_1 B + v'_2 B^2 + \cdots$$

The above transfer function is suitable because $|\delta'_1| < 1$ in order for the root of $(1 - \delta'_1 B) = 0$ to lie outside the unit circle and this in turn means that $(1 - \delta'_1 B)^{-1}$ causes the v'_i coefficients to decrease in absolute magnitude with increasing positive lags. A cyclic pattern in the residual CCF may indicate that $\delta'(B)$ should be at least second order in B.

(iv) The residual CCF from step (ii) can also be employed to ascertain the form of $\frac{\theta'(B)}{\phi'(B)}$ in [17.3.1]. As shown by Haugh and Box (1977), $\phi'(B) = \delta'(B)$ and $\theta'(B)$ should be at most of the order of $\delta'(B)$ or $\omega'(B)$.

(v) The results from steps (i) and (iii) can be substituted into [17.3.2] to obtain the form of $v(B)$. Likewise, the information from stages (i) and (iv) can be employed to get $\frac{\theta(B)}{\phi(B)}$ in [17.3.3]. The transfer function-noise model in [17.2.5] has now been completely identified.

Box and Jenkins Identification Procedure

By following the procedure of Box and Jenkins (1976) the model in [17.2.5] can be designed according to the following steps.

(i) Ascertain the most appropriate ARMA model in [16.2.3] to fit to the x_t series by utilizing the three stages of model construction presented in Part III. At the estimation stage, estimates are obtained for the ARMA model parameters and also the innovation series.

(ii) Using the ARMA filter, $\frac{\hat{\theta}_x(B)}{\hat{\phi}_x(B)}$, from step (i), transform the y_t series by employing

$$\hat{\beta}_t = \left[\frac{\hat{\theta}_x(B)}{\hat{\phi}_x(B)}\right]^{-1} y_t \qquad [17.3.5]$$

where the $\hat{\beta}_t$ sequence is usually not white noise since the filter in [16.2.4] is not used in [17.3.5].

(iii) After replacing the $\hat{v}_t$ series by the $\hat{\beta}_t$ sequence, use [16.2.6] to calculate the residual CCF for the $\hat{u}_t$ and $\hat{\beta}_t$ series.

(iv) Based upon the behaviour of the residual CCF from step (iii), identify the parameters required in the transfer function, $v(B)$, in [17.2.5]. As shown by Box and Jenkins (1976, p. 380), the theoretical CCF between the prewhitened input, u_t, and the correspondingly transformed output, β_t, is directly proportional to the impulse response function defined in Section 17.2.2. Consequently, the behaviour of the residual CCF can be utilized to identify the form of the transfer function in [17.2.5]. If, for instance, the residual CCF values are not significantly different from zero for negative lags, the estimated CCF at lag zero is significant, and the values attenuate for increasing positive lags, then the impulse response function must also follow this general behaviour. Therefore, an appropriate transfer function may be

$$v(B) = \frac{\omega_0}{1-\delta_1 B} = \omega_0(1+\delta_1 B+\delta_1 B^2+...)$$

$$= v_0 + v_1 B + v_2 B^2 + \cdots$$

since the v_i weights decrease in absolute magnitude with increasing positive lags due to the operator, $(1-\delta_1 B)$, with $|\delta_1| < 1$, in the denominator. A cyclic pattern in the residual CCF may imply that $\delta(B)$ should be at least second order in B. When the residual CCF possesses values which are significantly different from zero only at lags 0 and 1, the transfer function may be identified as

$$v(B) = \omega_0 - \omega_1 B = v_0 + v_1 B$$

(v) Subsequent to ascertaining the form of $v(B)$, determine the parameters needed in the noise term in [17.2.5]. Upon obtaining moment estimates for the parameters in $v(B)$, calculate the noise series from [17.2.5] by using

$$\hat{N}_t = (y_t - \bar{y}) - \hat{v}(B)(x_t - \bar{x})$$

where $\bar{y}$ and $\bar{x}$ are the sample means for μ_y and μ_x, respectively. By examining graphs such as the sample ACF and the sample PACF of $\hat{N}_t$, identify the ARMA model needed to fit to the noise series (see Section 5.3). Box and Jenkins (1976, pp. 384-385) also give a second procedure for identifying N_t where the sample CCF for $\hat{u}_t$ and $\hat{\beta}_t$, must first be calculated. The entire TFN model has now been tentatively designed.

Comparison of Identification Methods

The foregoing three identification procedures possess different inherent assets and liabilities. Although the empirical approach is straightforward and simple to apply, experience and understanding are required in order to properly identify the parameters needed in $v(B)$ at step (i). Because the empirical approach does not consider cross-correlation information in the first step, either the method of Haugh and Box or else Box and Jenkins could be utilized to check that $v(B)$ is properly designed. An advantage of the Haugh and Box method is that the residual CCF results that are employed for detecting causal relationships in Section 16.2.2 are also used for model identification. However, due to the relationships given in [17.3.2] and [17.3.3], the procedure is rather complicated and care should be taken that the model is not over-specified by having too many parameters. If this problem is not found at the identification stage, it may be detected at the estimation stage where some of the parameters may not be significantly different from zero. Common factors that appear in both the numerator and denominator of a transfer function should, of course, be left out of the model.

When using the Box and Jenkins approach, the form of $v(B)$ is identified directly from the CCF for the $\hat{u}_t$ and $\hat{\beta}_t$ series. However, in most cases $\hat{\beta}_t$ will not be white noise and therefore the estimated values of the CCF are correlated with one another (Bartlett, 1935). Consequently, caution should be exercised when examining the estimated CCF for $\hat{u}_t$ and $\hat{\beta}_t$.

In addition to the three identification methods presented in this section, other techniques are also available. For instance, Liu and Hanssens (1982) propose a procedure for identifying the transfer function parameters needed in $v(B)$ in [17.2.1] based upon least squares estimates of the transfer function weights using the original or filtered series. The corner method of Beguin et al. (1980) is then used to identify the order of the $\omega(B)$ and $\delta(B)$ operators. An advantage of the technique of Liu and Hanssens (1982) is that it is specifically designed for handling the situation where there are multiple input series to the TFN model as in [17.5.3]. A drawback of their technique is that it is fairly complicated to use in practice and, therefore, is not as convenient to employ as the empirical approach.

All of the identification techniques discussed in this chapter are designed for use in the time domain. An alternative approach to transfer function identification in the time domain is to identify a transfer function using frequency domain or spectral methods such as those suggested by Box and Jenkins (1976) and Priestley (1971). However, as noted by Liu and Hanssens (1982), spectral techniques are difficult to apply to practical problems.

When designing a TFN model, it is not necessary to adhere strictly to a given identification procedure. In certain situations, it may be advantageous to combine various steps from two or three of the three aforementioned identification methods which were discussed in detail. For example, either the method of Haugh and Box (1977) or the technique of Box and Jenkins (1976) could be employed to identify $v(B)$ in [17.2.5]. Step (ii) in the empirical approach could then be utilized to determine the form of N_t . In general, no matter what identification method is being utilized it is advantageous to begin with a fairly simple model, since the presence of too many parameters may cause the estimation procedure to become unstable. Because the results of the identification procedure can often be rather ambiguous, usually two or three possible models are suggested. If a suitable model is not included within the set of identified models, this will be detected at the estimation or diagnostic check stages of model construction. Either a more complicated model will be needed or further simplification will be possible due to having too many parameters.

17.3.2 Parameter Estimation

Following the identification of one or more plausible TFN models, efficient estimates must be simultaneously obtained for all of the model parameters along with their standard errors (SE's) of estimation. Because the a_t's are assumed to be normally, independently distributed, MLE's can be conveniently calculated for the model parameters along with their SE's. Appendix A17.1 explains how MLE's can be determined for the TFN model in [17.2.5]. As explained in that appendix, because the noise term in the TFN follows an ARMA process, one can expand an estimator developed for ARMA models for use in obtaining MLE's of the parameters in a TFN model. Moreover, an estimation procedure developed for use with a TFN model can also be employed with the intervention model of Chapters 19 and 22.

Often more than one tentative TFN model are initially identified. Subsequent to estimating the model parameters separately for each model, automatic selection criteria such as the AIC in [6.3.1] and the BIC in [6.3.5] can be utilized to assist in selecting the most appropriate model. Figure 6.3.1 outlines how the AIC or another appropriate automatic selection criterion can be incorporated into the three stages of model selection. If a suitable range of models is considered, it has been found in practice that the model possessing the minimum AIC value usually satisfies diagnostic tests of the model residuals. Nevertheless, the model or set of models that are thought to be most suitable should be thoroughly checked in order to ascertain if any further model improvements can be made.

17.3.3 Diagnostic Checking

The innovation sequence, a_t , is assumed to be independently distributed and a recommended procedure for checking the whiteness assumption is to examine a plot of the RACF (residual autocorrelation function) along with confidence limits. The RACF, $r_{\hat{a}\hat{a}}(k)$, can be calculated by replacing both $\hat{u}_t$ and $\hat{v}_t$ by $\hat{a}_t$ in [16.2.6] or else using [7.3.1]. Since $r_{\hat{a}\hat{a}}(k)$ is symmetric about lag zero, the RACF is only plotted against lags for $k = 1$ to $k \approx n/4$, along with the 95% confidence limits explained in Section 7.3.2. Although a plot of the RACF is the best whiteness test to use, other tests which can be employed include the cumulative periodogram in [2.6.2] and the modified Portmanteau test.

Three versions of the Portmanteau test for whiteness of the $\hat{a}_t$ residuals are given in Section 7.3.3 in [7.3.4] to [7.3.6]. In particular, the Portmanteau test in [7.3.6] is written as

$$Q_L = n\sum_{k=1}^{L} r_{\hat{a}\hat{a}}^2(k) + \frac{L(L+1)}{2n} \qquad [17.3.6]$$

where n is the number of data, $r_{\hat{a}\hat{a}}(k)$ is the residual CCF from [16.2.6] (replace both the $\hat{u}_t$ and $\hat{v}_t$ series by the $\hat{a}_t$ series in [16.2.6]), and L is a suitably chosen lag such that after L time periods a_t and a_{t-L} would not be expected to be correlated. For instance, when deseasonalized monthly data are being used in a TFN model, L should be chosen at least as large as lag 12 to make sure that there is no correlation between residuals which are separated by one year. Because Q_L is distributed as $\chi^2(L-p-q)$, where p and q are the orders of the AR and MA operators, respectively in the ARMA model for N_t , significance testing can be done to see if significant correlation of the model residuals is present.

If the residuals are correlated, this suggests some type of model inadequacy is present in the noise term or the transfer function, or both of these components. To ascertain the source of the error in the model, the CCF for the $\hat{u}_t$ and $\hat{a}_t$ sequences can be studied (leave $\hat{u}_t$ as $\hat{u}_t$ and replace $\hat{v}_t$ by $\hat{a}_t$ in [16.2.6] to estimate $r_{\hat{u}\hat{a}}(k)$). Because the u_t and a_t series are assumed to be independent of one another, the estimated values of $r_{\hat{u}\hat{a}}(k)$ should not be significantly different from zero where one standard error is approximately $n^{-1/2}$ when the CCF is normally distributed. When a plot of $r_{\hat{u}\hat{a}}(k)$ from $k \approx -n/4$ to $k \approx n/4$ along with chosen confidence limits indicate whiteness while significant correlations are present in $r_{\hat{a}\hat{a}}(k)$, the model inadequacy is probably in the noise term, $\hat{N}_t$. The form of the RACF for the $\hat{a}_t$ series could suggest appropriate modifications to the noise structure. However, if both $r_{\hat{a}\hat{a}}(k)$ and $r_{\hat{u}\hat{a}}(k)$ possess one or more significant values, where $r_{\hat{u}\hat{a}}(k)$ only has large values at non-negative lags, this could mean that the transfer function for the input series is incorrect and the noise term may or may not be suitable. When feedback is indicated by significant values of $r_{\hat{u}\hat{a}}(k)$ at negative lags, a multivariate model (see Part VII) should be considered rather than a TFN model.

Even though it is probably most informative to examine a plot of $r_{\hat{u}\hat{a}}(k)$ along with the 95% confidence limits, modified Portmanteau tests can also be employed to check if there are problems with the TFN model. To see whether or not $r_{\hat{u}\hat{a}}(k)$ has significantly large values at non-negative lags, the following modified Portmanteau statistic can be calculated.

$$Q_L = n^2 \sum_{k=0}^{L} r^2_{\hat{u}\hat{a}}(k)/(n-k) \qquad [17.3.7]$$

where Q_L is distributed as $\chi^2(L-r-m)$, and r and m are the orders of the $\delta(B)$ and $\omega(B)$ operators, respectively, in the transfer function for χ_t in [17.2.5]. If the calculated Q_L statistic is greater than the value of $\chi^2(L-r-m)$ from the tables at the chosen significance level, this could mean that the transfer function is incorrect and the noise term may or may not be suitable. By choosing more appropriate values of r and m, a model which passes this test can usually be found.

To check if $r_{\hat{u}\hat{a}}(k)$ has significantly large values at negative lags, the modified Portmanteau statistic can be determined using

$$Q_L = n^2 \sum_{k=-1}^{-L} r^2_{\hat{u}\hat{a}}(k)/(n+k) \qquad [17.3.8]$$

where Q_L is distributed as $\chi^2(L)$. If significance testing indicates that there are values of $r_{\hat{u}\hat{a}}(k)$ which are significantly different from zero at negative lags, this implies feedback and a multivariate model should be used (see Part IX) instead of a TFN model. Because $r_{\hat{u}\hat{a}}(-k) = r_{\hat{a}\hat{u}}(k)$, equation [17.3.8] can be equivalently written as

$$Q_L = n^2 \sum_{k=1}^{L} r^2_{\hat{a}\hat{u}}(k)/(n-k) \qquad [17.3.9]$$

Besides being independently distributed, the a_t sequence is assumed to follow a normal distribution and possess a constant variance (homoscedasticity). In Sections 7.4 and 7.5, tests are presented for checking the normality and homoscedastic suppositions, respectively. As noted in Section 3.4.5 as well as other parts of the book, in practice it has been found that a suitable Box-Cox transformation of the Y_t and/or X_t series can often correct non-normality and heteroscedasticity in the residuals.

Whenever problems arise in the model building process, suitable model modifications can be made from information at the diagnostic check and identification stages. Subsequent to estimating the model parameters for the new model, the modelling assumptions should be checked to see if further changes are necessary.

17.4 HYDROLOGICAL APPLICATIONS OF TRANSFER FUNCTION-NOISE MODELS WITH A SINGLE INPUT

17.4.1 Introduction

Two hydrological applications are presented for clearly explaining how TFN models are constructed in practice. In the first application, the residual CCF described in Section 16.2.2, is employed for determining the statistical relationship between monthly flows in the tributary of a river and the flows downstream in the main river. By using each of the three identification methods described in Section 17.3.1, tentative TFN models are designed for modelling the two monthly riverflow series and the most appropriate model is verified by diagnostic checking.

In the second application, a TFN model is designed for formally modelling one of the causal relationships discovered using the residual CCF in Section 16.2.3. The residual CCF studies from Chapter 16 can be considered as part of the *exploratory data analysis* stage where simple graphical and statistical tools are employed for detecting important statistical characteristics of the data (Tukey, 1977). At the *confirmatory data analysis* step, the TFN model in [17.2.5] can be utilized to formally model and confirm the mathematical relationships which are discovered at the exploratory data analysis stage. Accordingly, a dynamic model is developed for formally describing the connections between the annual Gota River flows and monthly temperatures.

17.4.2 Dynamic Model for the Average Monthly Flows of the Red Deer and South Saskatchewan Rivers

Identification

The South Saskatchewan (abbreviated as S.Sask.) River originates in the Rocky Mountains and flows eastward on the Canadian Prairies across the province of Alberta to Saskatchewan, where it joins the North Saskatchewan River northwest of the city of Saskatoon. These two rivers form the Saskatchewan River which flows into Lake Winnipeg in Manitoba, which in turn drains via the Nelson River into Hudson Bay. A major tributary of the S.Sask. River is the Red Deer River which connects to the S.Sask. River near the Alberta-Saskatchewan border. Average monthly flows in m^3/s are available from Environment Canada (1979a,b) for the Red Deer River near the city of Red Deer, Alberta and also for the S.Sask. River near Saskatoon, Saskatchewan. Saskatoon is located approximately 800 km downstream from the city of Red Deer and the area of the basin drained by the S.Sask. River at Saskatoon is 139,600 km^2, whereas an area of 11,450 km^2i, is drained by the Red Deer River at Red Deer.

Because the Red Deer River flows into the S.Sask. River it is obvious that the Red Deer River contributes to the overall flow of the S.Sask. River. However, even though the direction of causality can be easily physically justified a priori without a cross-correlation study, the results from a cross-correlation analysis can be employed to validate statistically the known causal relationship and also to design a TFN model that mathematically describes the dynamic connection between the input flows from the Red Deer River and output flows in the S.Sask. River.

Before obtaining the residual CCF, the flows must be prewhitened. When prewhitening a series it may be necessary to transform the data using the Box-Cox transformation given in [3.4.30]. Previously, hydrologists found by experience that a natural logarithmic transformation (i.e. $\lambda = 0$ in [3.4.30] and c = 0 when there are no zero flows) can preclude problems with heteroscedasticity and non-normality in the model residuals. Accordingly, for the time period from January, 1941, until December, 1962, the 264 values of the monthly flows for both the Red Deer and S.Sask. Rivers are transformed using natural logarithms. Subsequent to this, by using the deseasonalization procedure in [13.2.3], each time series is deseasonalized by subtracting out the monthly mean and dividing this by the monthly standard deviation for each data point in the logarithmic transformed data. The year 1962 is selected as the final year for which data are used because a large dam came into operation on the S.Sask. River after that time. By following the three stages of model construction given in Chapters 5 to 7, the most appropriate models from [16.2.3] and [16.2.4] to fit to the x_t and y_t series, respectively, are found to be ARMA(1,1) models. In Table 17.4.1, the MLE's and corresponding SE's for the model parameters are presented where the x_t series refers to the Red Deer flows and the y_t sequence represents the S.Sask. flows after taking natural logarithms and deseasonalizing the data.

Table 17.4.1. Parameter estimates for the Red Deer River and S.Sask. River ARMA(1,1) models.

RIVERS	PARAMETERS	MLE'S	SE'S
	$\phi_{x,1}$	0.845	0.045
Red Deer	$\theta_{x,1}$	0.292	0.080
x_t	σ_u^2	0.482	
	$\phi_{y,1}$	0.819	0.050
S.Sask.	$\theta_{y,1}$	0.253	0.084
y_t	σ_v^2	0.507	

The estimated white noise series, $\hat{u}_t$ and $\hat{v}_t$, for the x_t and y_t series, respectively, are automatically calculated at the estimation stage. By utilizing [16.2.6], the residual CCF in Figure 17.4.1 is calculated along with approximate 95% confidence limits. The large values at lags 0 and 1 statistically confirm the known physical fact that the Red Deer River causes flows in the S.Sask. River and not vice versa. As outlined in Table 16.2.1, because the residual CCF contains values which are only significantly different from zero at non-negative lags, there is unidirectional causality from X to Y. In Figure 17.4.1, the value of the residual CCF at lag -4 which just crosses the 95% confidence limits, can be attributed to chance.

As demonstrated by Figure 17.4.2, when the CCF for the x_t and y_t series are plotted the kind of causality cannot be statistically ascertained (note that the approximate 95% confidence limits in Figure 17.4.2 are derived for independent series). The large values at negative, zero and positive lags hide the known reality that the Red Deer River is a tributary of the S.Sask. River. Consequently, practitioners are urged to examine cautiously any CCF study where proper statistical procedures have not been followed.

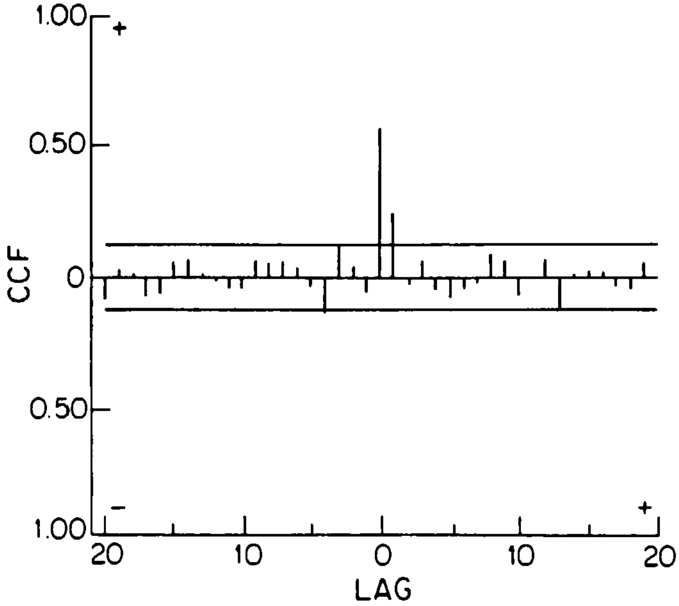


Figure 17.4.1. Residual CCF for the Red Deer and S.Sask. riverflows.

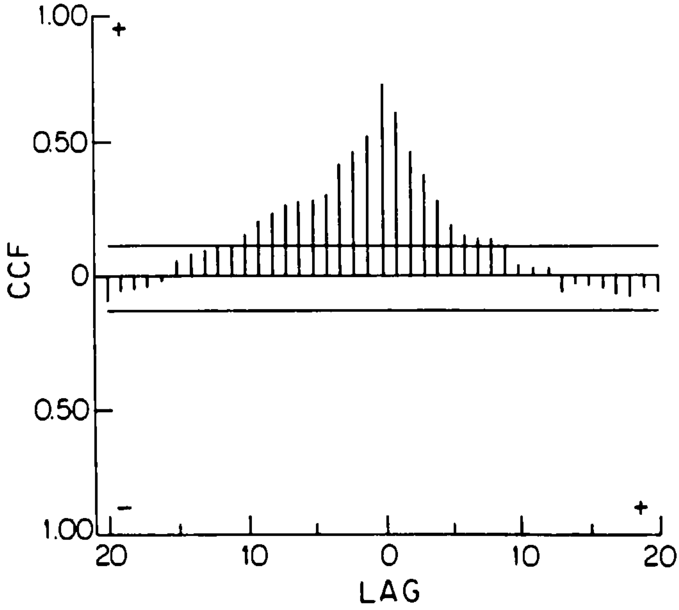


Figure 17.4.2. CCF for the deseasonalized logarithmic flows of the Red Deer and S.Sask. Rivers.

To explain how to design a TFN model for the x_t and y_t series for the Red Deer and S.Sask. Rivers, respectively, each of the three identification methods described in Section 17.3.1 is explained separately.

Empirical Identification Approach: Because monthly flows are being considered and also Saskatoon is about 800 km downstream from the city of Red Deer, from a physical point of view it would be expected that current Red Deer riverflows would affect the riverflows at Saskatoon during the present month and perhaps one month into the future. Consequently, a suitable transfer function may be

$$\nu(B) = \omega_0 - \omega_1 B$$

Assuming that the noise term in [17.2.5] is white noise, a TFN model is fitted to the x_t and y_t sequences, where the entries of the noise series are estimated along with the other model parameters. Using the standard model building procedures presented in Chapters 5 to 7, this series is found to be best described by an ARMA (1,1) model.

Haugh and Box Identification Method: Due to the large values at lags 0 and 1 of the residual CCF in Figure 17.4.1, the transfer function $\nu'(B)$ in [17.3.1] is identified to be

$$\nu'(B) = \omega'_0 - \omega'_1(B)$$

By employing [17.3.2] and also the results in Table 17.4.1, the transfer function in [17.2.5] which links x_t and y_t is calculated to be

$$\nu(B) = \frac{(1-0.253B)}{(1-0.819B)}(\omega'_0 - \omega'_1 B)\frac{(1-0.845B)}{(1-0.292B)}$$

$$\approx \omega'_0 - \omega'_1 B = \omega_0 - \omega_1 B$$

where the AR and MA operators can be dropped because $\theta_y(B) \approx \theta_x(B)$ and $\phi_y(B) \approx \phi_x(B)$ when the relative magnitudes of the standard errors are considered. Note that if it had been advantageous to get moment estimates for ω'_0 and ω'_1, [17.3.4] could have been utilized.

Since $\phi'(B) = \delta'(B)$, the order of the operator $\phi'(B)$ is zero. The order of $\theta'(B)$ should be at most of the order of $\delta'(B)$ or $\omega'(B)$ and, therefore, should be zero or one. Consequently, from [17.3.3] the noise term in [17.2.5] should be either ARMA (1,1) or else ARMA (1,2).

Box and Jenkins Procedure: By using [17.3.5], the $\hat{\beta}_t$ sequence is determined. The residual CCF for the $\hat{u}_t$ and $\hat{\beta}_t$ series is very similar to the plot in Figure 17.4.1 where there are large values only at lags 0 and 1. Accordingly, an appropriate transfer function for use in [17.2.5] is

$$\nu(B) = \omega_0 - \omega_1 B$$

Employing the same procedure used with the empirical identification method, the noise term is identified to be ARMA (1,1).

Parameter Estimation

The MLE's for the dynamic model linking x_t and y_t are listed in Table 17.4.2 where it is assumed that $\omega(B)$ is first order, $\delta(B)$ is of order zero, and the noise term is ARMA (1,1). The difference equation form of this model is written as

$$y_t = (0.572 + 0.238B)x_t + \frac{(1 - 0.494B)}{(1 - 0.856B)}a_t \qquad [17.4.1]$$

When the noise term is considered to be ARMA (1,2) the second MA parameter is not significantly different from zero and the value of the AIC is increased. Consequently, the simpler model in [17.4.1] is justified.

Table 17.4.2. Parameter estimates for the Red Deer-S.Sask. TFN model.

Parameters	MLE's	SE's
ω_0	0.572	0.049
ω_1	-0.238	0.049
ϕ_1	0.856	0.051
θ_1	0.494	0.085
σ_a^2	0.310	

Diagnostic Checking

The model in [17.4.1] satisfies the main modelling assumptions. In particular, the plot in Figure 17.4.3 of the RACF for the estimated a_t sequence along with 95% confidence limits reveals that the $\hat{a}_t$ series is white noise. The sample CCF and 95% confidence limits for the $\hat{u}_t$ and $\hat{a}_t$ sequences are displayed in Figure 17.4.4. Because the estimated values of the CCF and also the RACF in Figure 17.4.3 are not significantly different from zero, the transfer function and noise term are properly designed. Other diagnostic checks indicate that the $\hat{a}_t$ sequence is homoscedastic (see Section 7.5) and approximately normally distributed (see Section 7.4). Furthermore, the residual variance of 0.507 for the S.Sask. model in Table 17.4.1 is reduced by 39% to a value of 0.310 for the dynamic model in Table 17.4.2.

Concluding Remarks

Besides describing the dynamic relationship between the Red Deer and S.Sask. River, the model in [17.4.1] can be employed for applications such as forecasting and simulation. In fact, because the TFN model in Table 17.4.2 has a smaller residual variance than the ARMA (1,1) model in Table 17.4.1 for the S.Sask. River, it should produce more accurate forecasts. Forecasting with TFN models is explained and illustrated in Chapter 18.

17.4.3 Dynamic Model for the August Temperatures and Annual Flows of the Gota River

As presented in Table 16.3.3 for the causality studies of Section 16.3.3, the monthly temperature series are correlated with various annual riverflow series and the Beveridge wheat price index. The causality relationships which are found using the residual CCF as an exploratory data analysis tool can be further substantiated by developing a TFN model as a confirmatory data analysis tool. For illustrative purposes, one of the causal relationships in Table 16.3.3 is formally modelled here using a TFN model.

From Table 16.3.3, it can be seen that there are significant values of the residual CCF at lag zero between the flows of the Gota River and each of four monthly temperature series. This relationship is displayed graphically in Figure 16.3.3 for the case of the residual CCF for the August temperatures and the Gota riverflows. By following the model construction phases outlined in Section 17.3, an appropriate dynamic model can be developed (Hipel et al., 1985). When all four temperature series are used as covariate series in a TFN model, the transfer function parameter estimates for June, July and September are not significantly different from zero and can therefore be left out of the model. Whereas the residual CCF can only be used for pairwise comparisons, the TFN model can be employed to ascertain the most meaningful relationship when there are multiple covariate series and a single response or output series. For the case where only the August temperatures are used as a covariate series, the parameter estimates and SE's errors are listed in Table 17.4.3 while the difference equation for the model which follows the format of [17.2.5] is written as

$$Y_t - 535.464 = -23.869(X_t - 15.451) + (1 - 0.672B + 0.329B^2)^{-1}a_t \qquad [17.4.2]$$

where Y_t represents the annual flows of the Gota River and X_t stands for the monthly temperatures. Besides portraying the dynamic relationship between the Gota River flows and August temperatures, the model in [17.4.2] can, of course, be employed for forecasting and simulation.

17.5 TRANSFER FUNCTION-NOISE MODELS WITH MULTIPLE INPUTS

17.5.1 Introduction

In many situations, more than one input series is available for use in a TFN model and by incorporating all the relevant covariate series into the TFN model a dynamic model can be developed for producing more accurate forecasts and more realistic simulated values. For example, for the hydrometeorological application presented in Section 17.5.4, a flexible TFN model is constructed, where the average monthly precipitation and temperature series constitute the covariate series which affect the output or response consisting of average monthly riverflows. As demonstrated by this application, an inherent advantage of TFN modelling is that a TFN model with multiple inputs can be designed almost as easily as a model with a single input. In fact, the model building tools of Section 17.3 can easily be extended for use with a TFN model having multiple inputs.

In addition to using the most comprehensive statistical tools for use in model construction, the practitioner should exercise a lot of common sense and good judgement. As is the case for the other kinds of models considered in this text, TFN model building is in essence both an art and a science. The *art* of model building comes into play when the modeller uses his or her knowledge about the physical aspects of the problem to decide upon which covariate series should be incorporated into the TFN model and the general manner in which this should be done. For instance, for the application in Section 17.5.4, a suitable TFN model is developed by first considering simpler models which provide guidance as to how a more complex TFN model can be constructed. In the process of doing this, a simple procedure is suggested for creating a single input series which more than one precipitation or temperature series are available. By employing appropriate statistical and stochastic methods, the efficacy of the decisions made in the art of

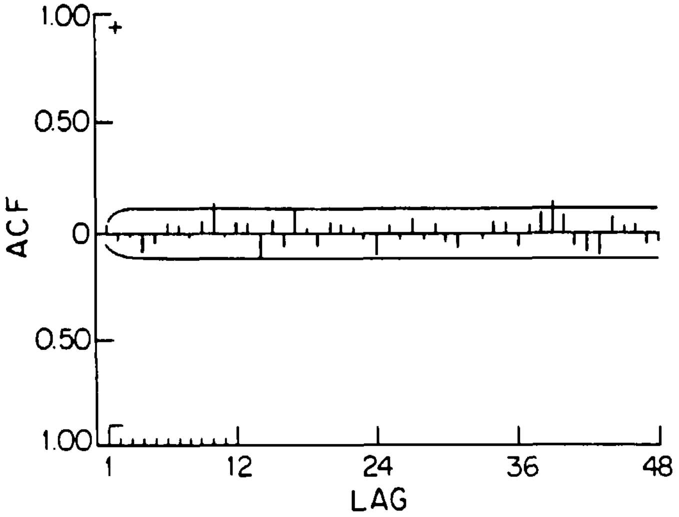


Figure 17.4.3. RACF for the Red Deer - S.Sask. TFN model.

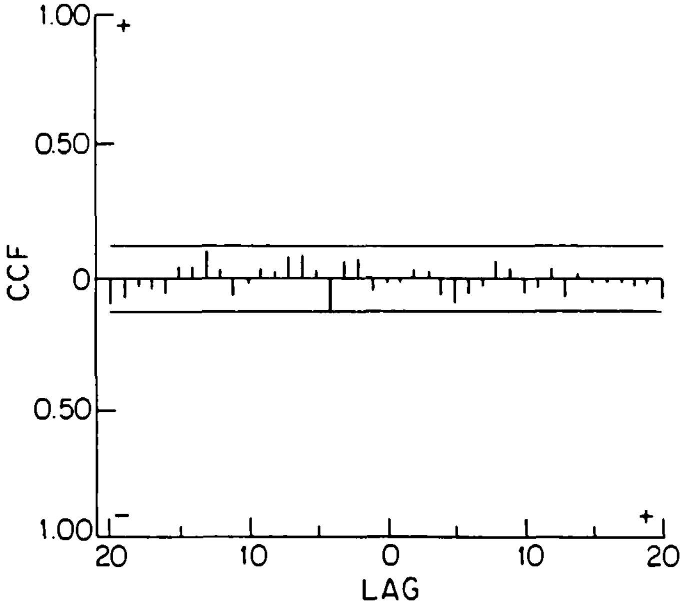


Figure 17.4.4. CCF of $\hat{u}_t$ and $\hat{a}_t$ for the Red Deer ARMA model and the Red Deer S.Sask. TFN model, respectively.

Table 17.4.3. Parameter estimates for the August temperature - Gota River flow transfer function-noise model.

PARAMETERS	MLE'S	SE'S
ω_0	-23.869	4.285
ϕ_1	0.672	0.077
ϕ_2	-0.329	0.077
σ_a	5.71×10^3	-

model building can be rigorously checked using the *science* of model construction. Consequently, there is interactive feedback when using both art and science for building TFN models, or for that matter, any other type of stochastic model.

17.5.2 Model Description

Qualitatively, the TFN model can be written as

output = dynamic component + noise

The dynamic component consists of a transfer function for each covariate series which describes how each input dynamically affects the output. As is also the case for a TFN model with one input in [17.2.5], the autocorrelated noise can be modelled using an ARMA model.

More precisely, a TFN model with multiple inputs can be written in the form

$$(y_t-\mu_y)=f(\mathbf{k},\mathbf{x},t)+N_t \qquad [17.5.1]$$

where t is discrete time, y_t is the output or response variable, μ_y is the mean of the y_t series, N_t is the stochastic noise term which may be autocorrelated, and $f(\mathbf{k},\mathbf{x},t)$ is the dynamic component of y_t . The dynamic component includes a set of parameters $\mathbf{k}$ and a group of covariate series $\mathbf{x}$. When required, both the response variable and one or more of the input variables may be transformed using a suitable Box-Cox transformation from [3.4.30]. As noted earlier, the reasons for transforming the series include stabilizing the variance and improving the normality assumption of the white noise series which is included in N_t . It should be pointed out that the same Box-Cox transformation need not be applied to all of the series. If the series are seasonal, subsequent to invoking appropriate Box-Cox transformations each series can be deseasonalized separately by employing the procedures of Section 13.2.2. Following this, identification procedures can be utilized to see which parameters should be included in the model in [17.5.1].

Included in the dynamic component of the model are the effects of all the input series upon the output. In general, if there are I input covariate series the dynamic component of the model is given by

$$f(\mathbf{k},\mathbf{x},t)=\sum_{i=1}^{I}v_i(B)(x_{ti}-\mu_{xi}) \qquad [17.5.2]$$

where x_{ti} is the ith input series which may be suitably transformed and μ_{xi} is the mean of x_{ti} . The ith transfer function which reflects the manner in which the ith input series, x_{ti} , affects y_t , is written as

$$\nu_i(B) = \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}$$

$$= \frac{(\omega_{0i} - \omega_{1i}B - \omega_{2i}B^2 - \cdots - \omega_{m_i i}B^{m_i})B^{b_i}}{(1 - \delta_{1i}B - \delta_{2i}B^2 - \cdots - \delta_{r_i}B^{r_i})}$$

where m_i and r_i are the orders of the operators $\omega_i(B)$ and $\delta_i(B)$, respectively, and b_i is the delay time required before x_{ti} affects y_t. Notice that the ith transfer function in [17.5.2] is identical to the one utilized in [17.2.1] and [17.2.5], except that the subscript i has been added to indicate that $\nu_i(B)$ is the transfer function for the ith input series x_{ti}.

In practice, usually only a few parameters are required in each transfer function and therefore m_i and r_i are 0 or 1 (see the applications in this chapter as well as Chapters 18, 19 and 22). Given the parameters for the $\omega_i(B)$ and $\delta_i(B)$ operators, it may be required to estimate the ν_{ji}, $j = 0,1,2,\ldots,$ coefficients in the operator

$$\nu_i(B) = (\nu_{0i} + \nu_{1i}B + \nu_{2i}B^2 + \cdots)$$

$$= \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}$$

These coefficients can be calculated in exactly the same manner as they are in Section 17.2.2 for the parameters of $\nu(B)$ which is the transfer function used for a TFN model with one input.

The noise component of the TFN model having multiple inputs is defined by

$$N_t = y_t - f(\mathbf{k},\mathbf{x},t)$$

That is, the noise term of the model is simply the difference between the response variable, y_t, and the dynamic component. The form of the noise term, N_t, is not restricted to any particular form, but usually it is assumed to be an ARMA process as in [17.2.4]. Furthermore, as noted in Section 17.2.3, the white noise component of the ARMA model is usually assumed to be $NID(0,\sigma_a^2)$. Finally, because the noise term models the portion of y_t which is not explained by the dynamic component, N_t is independent of each x_{ti} series. Equivalently, since the a_t disturbances drive N_t and each x_{ti} series can be thought of as being generated by its own residual series, the a_t sequence is independent of the white noise series for a given input series which can be formed by prewhitening the x_{ti} sequence (see Section 16.2.2 for a discussion of prewhitening). Finally, the x_{ti} series or their residuals formed by prewhitening are not assumed to be independent of one another in a TFN model.

In summary, by combining the dynamic and noise components, the overall TFN model with I inputs can be written as

$$(y_t - \mu_y) = f(\mathbf{k},\mathbf{x},t) + N_t \qquad [17.5.3]$$

$$= \sum_{i=1}^{I} \frac{\omega_i(B)}{\delta_i(B)} B^{b_i}(x_{ti} - \mu_{xi}) + \frac{\theta(B)}{\phi(B)} a_t$$

17.5.3 Model Construction

When developing a TFN model with multiple inputs, the parameters required in each transfer function within the overall dynamic component plus the orders of the operators in the noise term, must be identified. Subsequent to this, MLE's can be obtained for the model parameters and the validity of the model verified by invoking appropriate diagnostic checks. Because the TFN model with multiple inputs in [17.5.3] is a straightforward extension of the single input model in [17.2.5], most of the construction tools presented in Section 17.3 for a TFN having one input can be used for the multiple input case. The purpose of this section is to clearly point out what special problems can arise when building a TFN model with more than one input and how a modeller should cautiously use the identification tools of Section 17.3.1 for the multiple input case.

In Section 17.3.1, the following three identification procedures are explained in detail for the case of a TFN model with a single input:

(i) the empirical identification approach,

(ii) the Haugh and Box identification method, and

(iii) the Box and Jenkins identification procedure.

All three methods were specially developed under the assumption that only one input series is present in the model and the input series only affects the output. In fact, this assumption is theoretically embedded into the latter two procedures. When there is more than one input series, the obvious way to use each identification procedure is to investigate, pairwise, the relationship between each x_{ti} series and the y_t in order to design the form of the transfer function $v_i(B)$. However, in a TFN model with more than one covariate series, two or more covariates may not be independent of one another and may therefore affect each other in addition to driving the response variable. If there is not too much interaction among the x_{ti} series, fairly correct transfer functions may be identified using the pairwise identification procedure. Whatever the case, the assumption that the x_{ti} 's are independent of one another is not assumed for the TFN model itself in [17.5.3]. Therefore, if required, a number of tentative models can be fitted to the series. After also identifying the noise term, different discrimination techniques can be used to isolate the most appropriate model or set of models. For example, one can select the model having the lowest value of the AIC in [6.3.1] or the BIC in [6.3.5]. One can also remove any parameter from a model whose estimated value is not significantly different from zero. Finally, one should also insure that this model passes diagnostic checks, especially the tests for determining the whiteness of the estimated a_t series.

When designing the transfer function, the most suitable approach is probably to use the empirical approach in conjunction with the method of Haugh and Box. Although it may be complicated to use in practice, another procedure for identifying transfer functions is to employ the method of Liu and Hanssens (1982) which is specifically designed for identifying TFN models with multiple inputs. However, when deciding upon the parameters needed in the noise term, N_t, it is recommended for most applications that the empirical technique be used. Recall from Section 17.3.1 for the empirical approach, that after designing the transfer functions, the model in [17.5.3] is fitted to the set of series where it is initially assumed the noise term is white, even though it is probably not. Next, by following the model development phases of Chapters 5 to 7 for a single series, the best ARMA model is selected for modelling the estimated noise sequence form the previous step. Keep in mind, that because the noise term is assumed to be independent of the dynamic component, this is a theoretically valid procedure. Finally, the identified noise term can be used in [17.5.3] and then all of the model parameters can be simultaneously estimated for the completely identified TFN model.

Because the a_t's are assumed to be $NID(0,\sigma_a^2)$, MLE's can be efficiently calculated using the method of McLeod (1977) or another appropriate estimate such as one of those listed in Section 6.2.3. Not only are the MLE's obtained using McLeod's method almost exact MLE's, but the computation time required is much lower than that needed by other available exact MLE procedures. The estimation procedure described in Appendix A17.1 for a TFN with one input, can easily be expanded for use with a TFN model having multiple inputs.

At the model validation stage, the key assumption to check is that the residuals, $\hat{a}_t$, which are estimated along with the model parameters, are white. As explained in Section 17.3.3, this can be accomplished by investigating a plot of $r_{\hat{a}\hat{a}}(k)$ from [16.2.6] or [7.3.1] along with the 95% confidence limits. If there are problems, the form of the significantly large autocorrelations present in $r_{\hat{a}\hat{a}}(k)$ may indicate what type of model modifications should be made to either N_t, the dynamic component, or both. Investigating the form of the residual CCF between each prewhitened x_{ti} series and $\hat{a}_t$ may also assist in detecting where the sources of the problems are located and how they should be rectified. However, if the x_{ti} series were not previously prewhitened for use in a causality study (see Section 16.2.2) or some other purpose, obtaining the residual CCF pairwise for each prewhitened x_{ti} series and $\hat{a}_t$ may be quite time consuming. Furthermore, the alterations suggested by each individual residual CCF may not necessarily hold for the overall TFN model in [17.5.3] because of possible interactions among the x_{ti} series themselves. Fortunately, the authors have found in practice that when the empirical identification approach is utilized in conjunction with a sound understanding of the physical realities of the problem being studied, usually problems with the design of the TFN model can be circumvented.

17.5.4 Hydrometeorological Application

Introduction

The general procedure in many modelling problems is to start with a simple model and then increase the model complexity until an acceptable description of the phenomenon is achieved or until further improvements in the model cannot be obtained by increasing the model complexity. This is especially true in TFN modelling where there is a single response variable and multiple input series. However, the question arises as to how one can conveniently construct the most effective model for describing the dynamic relationships between the output and input series.

The purpose of the hydrometeorological application in this section is to demonstrate the *art and science of building a suitable TFN model.* As will be shown, a range of different TFN models are developed and the most appropriate model is systematically found. The output for each TFN model always represents the deseasonalized average monthly logarithmic flows of the Saugeen River at Walkerton, Ontario, Canada, while the covariate series consist of either transformed precipitation or temperature data sets, or both types of series. The types, lengths and locations of measurement for the data sets entertained are shown in Table 7.5.1 for the single riverflow sequence, the two precipitation and the two temperature series. The riverflow data are obtained from Environment Canada (1980a) and the precipitation and temperature data are provided by the Atmospheric Environment Service in Downsview Ontario (Environment Canada, 1980b).

Table 17.5.1. Available monthly data.

Type	Location	Period
Riverflows	Saugeen River at Walkerton, Ontario	1963-1979
Precipitation	Paisley, Ontario	1963-1979
Precipitation	Lucknow, Ontario	1950-1979
Temperature	Paisley, Ontario	1963-1979
Temperature	Lucknow, Ontario	1950-1979

When the data are used in the upcoming application as series in the CCF analyses or as input or output series in the TFN models, the time series are only employed for the time period during which all the series overlap. However, when estimating missing observations within a single time series, the entire time series is used (see Section 19.3). For a further discussion and the original presentation of this application, the reader may wish to refer to the paper of Hipel et al. (1982).

Missing Data

Prior to constructing a TFN model, any missing data in the covariate series must be estimated. The only missing data in this study are ten precipitation and corresponding temperature data points for the Lucknow station where the dates of these missing data are given in Table 16.5.2. As is explained in detail in Section 19.3 in the chapter on intervention analysis, a special type of intervention model can be designed for obtaining good estimates of the missing data points. An inherent advantage of the intervention analysis approach for estimating missing data points is that the correlation structure of the series is automatically taken into account when obtaining the estimates for the missing observations. When the intervention model in [19.3.7], developed in Section 19.3.6 for the entire deseasonalized Lucknow temperature data, is employed for data filling, the estimates and SE's given in the third column of Table 17.5.2 are obtained for the original series. The estimates for the missing observations can be compared to their respective monthly means in the second column of Table 17.5.2. For this particular application, the difference between each estimate and its monthly mean is always less than its SE.

Table 17.5.2. Estimates of missing temperature data at Lucknow.

Dates	Monthly Means (C°)	Estimates (C°) (SE's)
February 1953	-6.48	-6.57 (2.32)
May 1968	11.99	11.81 (1.78)
September 1968	15.17	15.34 (1.16)
October 1973	9.60	9.78 (1.67)
August 1975	18.89	18.59 (1.20)
September 1975	15.17	15.29 (1.16)
July 1976	19.68	19.47 (1.09)
September 1978	15.17	15.30 (1.16)
October 1978	9.60	8.30 (1.70)
August 1979	18.89	18.56 (1.18)

When the Lucknow precipitation data is deseasonalized using [13.2.3], the resulting nonseasonal sequence is white noise. Because there is no correlation structure in the series, the appropriate estimate of each missing data point is simply taken as its average monthly value.

The intervention analysis approach to data filling in Section 19.3 can be used when not more than about 10% of the data are missing. If there are many missing observations, where there may be rather long periods of time over which there are no data at all, the technique of Section 22.2 may be useful. Furthermore, when there are two series which are causally related but one is longer than the other, a TFN model relating the two series can be used to obtain estimates of the shorter series where it doesn't overlap with the longer one. This technique is called backforecasting and is explained in Section 18.5.2. Whatever the case, once the unknown observations have been estimated for each input series, a TFN model can be built for describing the relationships between the output and the inputs.

Identifying the Dynamic Component

Based upon a physical understanding of the problem and also using residual CCF analyses, the possible forms of the transfer functions can be identified for linking precipitation or temperature to the riverflow output. Firstly, the Saugeen River flows are transformed using a logarithmic transformation ($\lambda = 0$ in [3.4.30]) in order to avoid problems of non-normality and/or heteroscedasticity in the model residuals. It is found that the precipitation and temperature series do not require a power transformation. Next, an ARMA model is fitted to each deseasonalized series in Table 17.5.1 and the model residuals are estimated. Finally, the residual CCF between the residuals from the model fitted to the Saugeen River flows and each of the other four residual series are then calculated using [16.2.6].

The results of the cross-correlation analyses show a positive significant relationship at lag zero for each of the two precipitation series. For instance, the plot of the CCF for the Lucknow precipitation and Saugeen riverflows is displayed in Figure 17.5.1 along with the estimated 95% confidence interval. The value of the residual CCF at lag zero in Figure 17.5.1 is 0.448 whereas for the Paisley precipitation the estimated value of 0.365 is slightly smaller. Although the residual CCF plot for the Paisley precipitation is not shown, it is indeed similar in form to Figure

17.5.1. The characteristics of the residual CCF's for the two precipitation series makes intuitive sense from a physical point of view since for monthly data, most of the precipitation for a particular month will result in direct runoff in the same month.

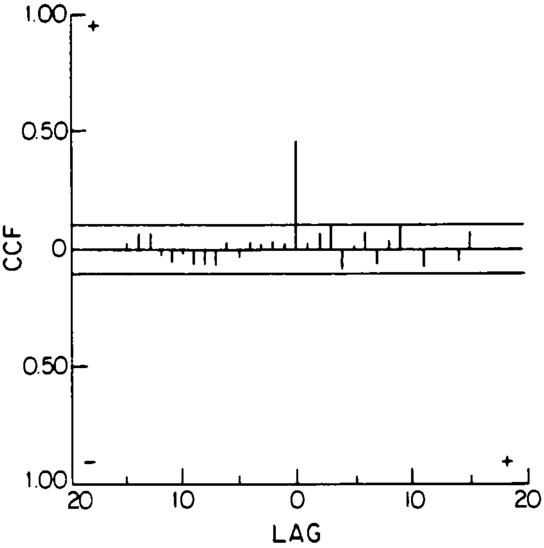


Figure 17.5.1. Residual CCF for the Lucknow precipitation and Saugeen riverflows.

The results of the residual CCF analyses for the two temperature series and the Saugeen riverflows are somewhat different. In these cases, there are no significant cross-correlations at any lag. However, from a physical viewpoint, one might expect that above average temperatures during the winter season would increase snowmelt and thus riverflow. For this reason, the temperature series are considered in the upcoming TFN model building. The temperature series are assumed to have a significant contribution at lag zero and as will be shown later, this assumption is found to be justifiable.

Combining Multiple Times Series

Often more than one covariate series of a particular kind is available to the analyst. In hydrological studies, data from several precipitation and temperature stations within or near the basin may be available. A common procedure employed by hydrologists to reduce model complexity is to combine similar types of series to form a single input covariate series. In the case of precipitation data, the records from the various stations are often combined to provide a single series of mean precipitation for a given region or basin. Two common methods of combining precipitation series are the Isohyetal and the Thiessen polygon techniques (Viessman et al., 1977). These procedures are essentially graphical methods and require a skilled analyst to obtain reasonable and consistent results. In an effort to automate procedures for combining similar types of series and provide more consistent results, a technique based on combining transfer function coefficients is presented.

Consider the case where two input covariate series, x_{t1} and x_{t2}, are to be combined to form a single input covariate series x_t. If x_{ti} causes y_t instantaneously, then the TFN models for the two series would be

$$y_t = \omega_{01}x_{t1} + N_{t1} \qquad [17.5.4]$$

and

$$y_t = \omega_{02}x_{t2} + N_{t2} \qquad [17.5.5]$$

where ω_{01} and ω_{02} are the transfer function parameters for the series x_{t1} and x_{t2}, respectively. For this example, the two series x_{t1} and x_{t2} would be combined using the relative ratio of the transfer function coefficients such that

$$\left(\frac{\omega_{01}}{\omega_{01}+\omega_{02}}\right)x_{t1} + \left(\frac{\omega_{02}}{\omega_{01}+\omega_{02}}\right)x_{t2} = x_t \qquad [17.5.6]$$

If more than two input series were available, this procedure could simply be extended to combine all of the available data into one input covariate series.

The Transfer Function-Noise Models

The various TFN models and one ARMA model that are examined, as well as their associated AIC values, are presented in Table 17.5.3. A decrease in the value of the AIC indicates that the additional model complexity is probably warranted since a better statistical fit is obtained. As expected, each increase in model complexity leads to a corresponding decrease in the value of the AIC. Thus a better description of the phenomena results with each addition of available information. For all of the TFN models in Table 17.5.3, the noise term is identified using the empirical approach of Section 17.3.1 to be an ARMA(1,0) model. Furthermore, all of the models satisfy the diagnostic checks presented in Section 17.3.3. Details of each of the models are now discussed.

The first model considers only the Saugeen riverflows. The time series is first transformed by taking natural logarithms of the data. This series is then deseasonalized by subtracting the estimated monthly mean and dividing by the estimated monthly standard deviation for each observation as in [13.2.3]. An ARMA(1,0) model is found to be the best model to fit to these deseasonalized flows. The value of the AIC is 963.121 and this value is used as a basis for comparing improvements in each of the subsequent models in Table 17.5.3.

Precipitation Series as Inputs: As suggested by the results of the residual CCF analyses, each of the precipitation series is used as an input covariate series. Prior to fitting the TFN model, each of the series is first deseasonalized. Each series is then used independently as an input covariate series in a TFN model. As shown in Table 17.5.3, the transfer function parameter, ω_0, for Paisley and Lucknow are estimated as 0.310 and 0.350, respectively. Note that the AIC value for the Lucknow precipitation series is significantly less than the AIC value for Paisley. This may suggest that the pattern of the overall precipitation which falls on the Saugeen River basin upstream from Walkerton, is more similar to the precipitation at Lucknow than the precipitation at Paisley, even though Paisley is closer to Walkerton than Lucknow.

Table 17.5.3. Transfer function-noise models fitted to the data.

Output and Covariate Time Series	Parameter Estimates (SE's)	AIC
Saugeen Flows	$\phi_1 = 0.407\ (0.064)$	963.121
Saugeen Flows Paisley Precipitation (ω_0)	$\hat{\phi}_1 = 0.405\ (0.064)$ $\hat{\omega}_0 = 0.3/0\ (0.055)$	936.129
Saugeen Flows Lucknow Precipitation (ω_0)	$\hat{\phi}_1 = 0.429\ (0.063)$ $\hat{\omega}_0 = 0.350\ (0.053)$	925.606
Saugeen Flows Combined Precipitation (ω_0)	$\hat{\phi}_1 = 0.418\ (0.064)$ $\hat{\omega}_0 = 0.350\ (0.054)$	926.340
Saugeen Flows Summer Precipitation (ω_{01}) Winter Precipitation (ω_{01})	$\hat{\phi}_1 = 0.407\ (0.064)$ $\hat{\omega}_{01} = 0.444\ (0.065)$ $\hat{\omega}_{02} = 0.183\ (0.092)$	922.270
Saugeen Flows Summer Precipitation (ω_{01}) Accumulated Snow (ω_{02})	$\hat{\phi}_1 = 0.408\ (0.064)$ $\hat{\omega}_{01} = 0.448\ (0.066)$ $\hat{\omega}_{02} = -0.162\ (0.109)$	924.037
Saugeen Flows Paisley Temperature (ω_0)	$\hat{\phi}_1 = 0.414\ (0.064)$ $\hat{\omega}_0 = 0.073\ (0.061)$	963.700
Saugeen Flows Lucknow Temperature (ω_0)	$\hat{\phi}_1 = 0.419\ (0.064)$ $\hat{\omega}_0 = 0.112\ (0.062)$	961.860
Saugeen Flows Summer Precipitation (ω_{01}) Winter Precipitation (ω_{02}) Lucknow Temperature (ω_{03})	$\hat{\phi}_1 = 0.422\ (0.064)$ $\hat{\omega}_{01} = 0.453\ (0.064)$ $\hat{\omega}_{02} = 0.194\ (0.090)$ $\hat{\omega}_{03} = 0.144\ (0.055)$	917.558
Saugeen Flows Summer Precipitation (ω_{01}) Winter Precipitation (ω_2) Combined Temperature (ω_{03})	$\hat{\phi}_1 = 0.420\ (0.064)$ $\hat{\omega}_{01} = 0.453\ (0.064)$ $\hat{\omega}_{02} = 0.194\ (0.090)$ $\hat{\omega}_{03} = 0.133\ (0.055)$	918.586

Using the procedure outlined in the previous section, the two precipitation series are combined to form a single input covariate series. In this study, the Lucknow and Paisley precipitation series are combined in the ratio 53:47, respectively. This combined precipitation series is then deseasonalized and used as an input series for the TFN model. The resulting AIC value is only slightly larger than the AIC value obtained when only the Lucknow precipitation series is employed. Since the difference is small, either model would be satisfactory and for the balance of this section the combined precipitation series is employed.

In the previous models, the precipitation series are entered as a single series having a single transfer function parameter. In these cases, it is therefore assumed that the contribution of precipitation is the same throughout the year. Physically, however, it makes sense that the contribution of precipitation during the winter months would be less than the contribution during the warmer periods of the year, since the precipitation accumulates on the ground in the form of snow during the cold season. In an effort to better reflect reality, the single precipitation series formed by combining the Lucknow and Paisley data, is divided into two separate seasons.

In separating the precipitation into two seasons, the winter season is taken as those months where the mean monthly temperature is below zero degrees Celcius. For both the Paisley and the Lucknow temperature series, December, January, February and March have mean monthly temperatures below freezing. Therefore, the winter precipitation series consists of the deseasonalized precipitations for these four months and zeros for the other eight months of the year. Conversely, the summer precipitation series has zeros for the four winter months and the deseasonalized precipitations for the remaining entries. These two series are input as separate covariate series with separate transfer function parameters. The resulting calibrated TFN model is

$$y_t = 0.444x_{t1} + 0.183x_{t2} + \frac{a_t}{(1-0.407B)} \qquad [17.5.7]$$

where y_t is the deseasonalized logarithmic riverflow at time t; x_{t1} is the combined deseasonalized summer precipitation series; and x_{t2} is the combined deseasonalized winter precipitation series. As expected, the transfer function coefficient for the summer precipitation is larger than the transfer function parameter for the winter precipitation. It is also reassuring to note that the better representation of the physics of the system also leads to an improved statistical fit as indicated by a lower AIC value.

A second type of dynamic model aimed at modelling the spring runoff resulting from snowmelt is also considered. In this model, the summer precipitation is the same as the model in [17.5.7]. However, the snow accumulated during the winter months from December to March is represented as a single pulse input in April where the temperature is above zero for the first time and hence spring runoff occurs. For the other eleven months of the year this series has zero entries. This type of dynamic model has been shown to work well for river systems located in areas that experience Arctic climate (Baracos et al., 1981) and rarely have any thaws during the winter months. However, the climatic conditions in the Saugeen River basin during the winter are not extremely cold and several midwinter melts result in a significant reduction in the accumulated snow cover on the ground. For this reason, the transfer function parameter for the accumulated winter precipitation is not significantly different from zero. Notice in Table 17.5.3 that that value of $\hat{\omega}_{02} = -0.162$ is much less than twice the SE of 0.109.

When dealing with quarter monthly data, another approach is presented in Section 18.3 and also by Thompstone (1983, Ch. 6) for incorporating snowmelt into a TFN model. The advantage of this approach is that it closely reflects the physical processes of snowmelt.

Temperature Series as Inputs: Although the residual CCF analyses indicate no significant relationships between temperature and riverflow, the two temperature series are used as input covariate series in TFN models. As before, both series are first deseasonalized by subtracting out the estimated monthly means and dividing by the estimated monthly standard deviations for each observation as in [13.2.3]. These series are then entered independently as covariate series in

TFN models. The resulting models and their associated AIC values are shown in Table 17.5.3. Because 1.96 times the SE for each parameter is larger than the parameter estimate, neither transfer function parameter is significant at the five percent significance level. Recall that the CCF for each temperature series and the Saugeen riverflows, also suggests that there may not be a marked relationship between the temperatures and riverflows. However, the Lucknow temperature parameter is significantly different from zero at the ten percent significance level. As a result, the Lucknow temperature series is included in the TFN models where both the temperature and precipitation series are included.

Precipitation and Temperature Series as Inputs: In an effort to combine all of the available information, both the temperature and the precipitation data are used as input covariate series in TFN models. In the first model of this type in Table 17.5.3, the combined precipitation is deseasonalized and split into two seasons as is done in [17.5.7]. The deseasonalized Lucknow temperature data is used as another input covariate series. The resulting model is given by

$$y_t = 0.453x_{t1} + 0.194x_{t2} + 0.144x_{t3} + \frac{a_t}{(1 - 0.422B)} \qquad [17.5.8]$$

where y_t is the deseasonalized logarithmic flow at time t; x_{t1} is the deseasonalized summer precipitation; x_{t2} is the deseasonalized winter precipitation; x_{t3} is the deseasonalized Lucknow temperature data; and a_t is the white noise term. The model and its associated AIC are also shown in Table 17.5.3. This model provides a significant improvement over any of the models previously employed with a decrease of almost five in the AIC when compared to the model in [17.5.7]. Also, the transfer function parameter for the temperature series is significantly different from zero in this case. Recall from before that the transfer function parameter for either temperature series is not significantly different from zero at the five percent significance level. However, when the precipitation series is included in the model, the temperature series provides a significant contribution. This point illustrates the need for more research in identifying the dynamic component of TFN models when more than one input covariate series is available.

The last model fitted to the data employs the combined precipitation and the combined temperature data. The temperature series are combined in the same fashion as the precipitation series but are not divided into two separate seasons. The resulting model and its associated AIC are shown in Table 17.5.3. Note that the AIC value is only marginally larger than that of the previous model where the Lucknow temperature data is employed instead of the combined temperature series. In this case, either of these last two TFN models could be employed as the most appropriate model for the available data.

17.6 ARMAX MODELS

From [17.2.7], the TFN model having a single covariate series is written as

$$y_t - \mu_y = \frac{\omega(B)}{\delta(B)}B^b(x_t - \mu_x) + \frac{\theta(B)}{\phi(B)}a_t \qquad [17.6.1]$$

The first and second terms on the right hand side of [17.6.1] are referred to as the dynamic and noise components in Sections 17.2.2 and 17.2.3, respectively. The operators contained in the transfer function $\frac{\omega(B)}{\delta(B)}B^b$ for transferring the influence of the covariate series x_t to the response series y_t are defined in Section 17.2.2. Finally, the operators in the transfer function $\frac{\theta(B)}{\phi(B)}$ for the white noise a_t are defined in [17.2.4] of Section 17.2.3 for the ARMA process describing the correlated or coloured noise as

$$N_t = \frac{\theta(B)}{\phi(B)}a_t \qquad [17.6.2]$$

Some authors prefer to write a dynamic model having a stochastic noise component in a different fashion than the one given in [17.6.1]. More specifically, their model is written as

$$\phi(B)(y_t - \mu_y) = \omega(B)B^b(x_t - \mu_x) + \theta(B)a_t \qquad [17.6.3]$$

In the literature, this model is most commonly called the *ARMAX (autoregressive-moving average-exogeneous variables) model* (Hannan, 1970) but is also referred to as the ARTF (autoregressive transfer function) and ARMAV models. The exogeneous variable in [17.6.3] refers to the input series, x_t. By dividing [17.6.3] by $\phi(B)$, the ARMAX model can be equivalently given as

$$(y_t - \mu_y) = \frac{\omega(B)}{\phi(B)}B^b(x_t - \mu_x) + \frac{\theta(B)}{\phi(B)}a_t \qquad [17.6.4]$$

A feature of the ARMAX model written in [17.6.3] is that the response variable is written directly as an autoregression. For example, if the order of $\phi(B)$ is 2, the model would be

$$y_t - \mu_y = \phi_1(y_{t-x} - \mu_h) + \phi_2(y_{t-2} - \mu_y) + \omega(B)B^b(x_t - \mu_x) + \theta(B)a_t$$

Nonetheless, the TFN model possesses distinct theoretical and practical advantages over the ARMAX model. To compare these two models consider the linear filter interpretations depicted in Figures 17.6.1 and 17.6.2 for the TFN model in [17.4.1] and ARMAX model in [17.6.4], respectively. In each figure, the inputs to the model are the a_t innovations plus the covariate x_t series. After passing through the indicated linear filters, the plus sign indicates that the dynamic and noise components are added together to create the y_t response. In both models, the input signal, x_t, and white noise, a_t, enter the linear filter system by different pathways. However, for the case of the ARMAX model in Figure 17.6.2, the input transfer function $\frac{\omega(B)}{\phi(B)}$ and the ARMA noise transfer function $\frac{\theta(B)}{\phi(B)}$ are interrelated, since they have the same common denominator $\phi(B)$. On the other hand, the transfer functions for the input and noise terms for the TFN model in Figure 17.6.1 have no common operators and are, therefore, independent of one another. Consequently, the TFN model is a more general representation of a dynamic-noise system than the ARMAX model (Young, 1984, pp. 113-116). The TFN model clearly separates out the deterministic or dynamic component and the stochastic noise effects.

As pointed out by Young (1984, pp. 116-117), one could also define other related versions of a dynamic-noise model outside of those given in [17.4.1] and [17.6.3]. For example, one could constrain the noise term in [17.6.1] or [17.6.4] to be purely AR or solely MA. Overall, the most general and flexible definition is the TFN model given in [17.6.1].

A great practical advantage of TFN modelling is that comprehensive model construction tools are available for conveniently applying TFN models to realworld problems. As explained in Section 17.3, flexible techniques are known and tested for use in identification, estimation and diagnostic checking of TFN models. However, this is not the situation for ARMAX models.

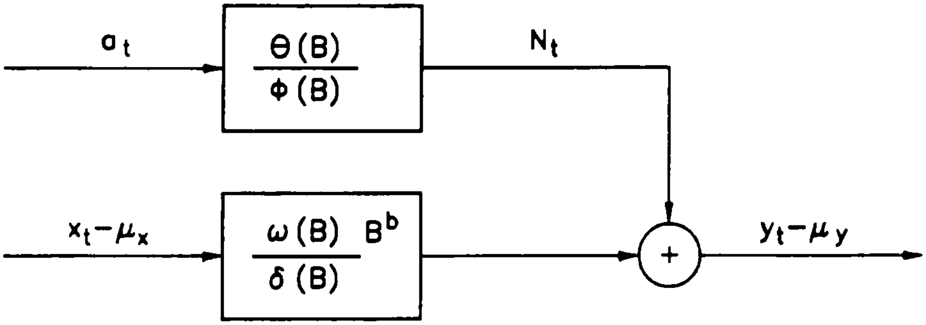

a_t
$\frac{\theta(B)}{\phi(B)}$
N_t
$x_t-\mu_x$
$\frac{\omega(B)\ B^b}{\delta(B)}$
+
$y_t-\mu_y$

Figure 17.6.1. Linear filter depiction of the TFN model in [17.6.1].

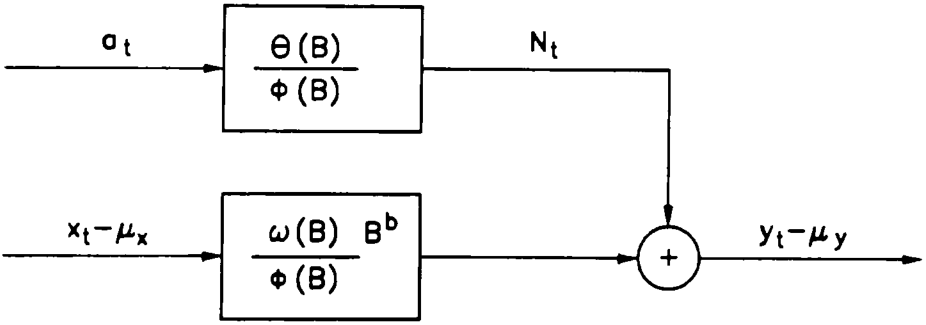


Figure 17.6.2. Linear filter interpretation of the ARMAX model in [17.6.4].

For example, one has to be careful when estimating the parameters of an ARMAX model. If the order of certain operators in the ARMAX model are not specified correctly, one may obtain poor parameter estimates (Young, 1984, p. 117). Because of the aforesaid and other reasons, it is recommended that the TFN model be selected over the ARMAX model for employment in practical applications.

The ARMAX model was introduced into system identification by Astrom and Bohlin (1965). In the past, ARMAX models have been successfully applied to hydrological and environmental engineering problems. For documented case studies of ARMAX modelling in hydrology, the reader can refer, for instance, to the research of Haltiner and Salas (1988), as well as references cited therein. In environmental engineering, ARMAX models have been employed for modelling dynamic systems problems arising in the conveyance (Capodoglio et al., 1990) and treatment (Capodoglio et al., 1991, 1992; Novotny et al., 1991) of wastewater. As a matter of fact, ARMAX and TFN models perform well when compared to more complex deterministic models written as differential equations, for capturing the key dynamic aspects of wastewater treatment plants (Capodoglio et al., 1992).

17.7 CONCLUSIONS

As exemplified by the applications in this chapter, a TFN model can be conveniently constructed for handling both single and multiple inputs. In addition to an understanding of the physical properties of a system being modelled, an array of well developed statistical tools are available for use in designing a suitable TFN model. For model identification, the empirical approach along with appropriate CCF analyses described in Section 17.3.1, can be used to identify the dynamic component of a TFN model. To identify the noise component of the TFN model, it is recommended that the empirical approach be used, especially when there are more than one covariate series. Subsequent to identifying one or more tentative TFN models, MLE's can be obtained for the model parameters (see Appendix A17.1) and a number of statistical tests can be employed for checking the validity of the model. Automatic selection criteria, such as the AIC and BIC, can be quite useful for model discrimination purposes.

If there are more than one precipitation or temperature series, a procedure is available for obtaining a single precipitation or temperature series. For the situation where snow accumulates during the winter time, the precipitation series can be incorporated into the dynamic model in specified manners so that the model makes sense from a physical point of view. For the case of the best Saugeen River dynamic model, the precipitation series was split into a winter and summer series, and a separate transfer function was designed for each of the series. An alternative approach for incorporating snowmelt into a TFN model is presented in Section 18.3.

When TFN models are fitted to other kinds of environmental series, the scientist can practice the *art and science* of his profession by designing physically based transfer functions and using flexible statistical tools to isolate the best design. As shown by the TFN applications in the next chapter, a properly designed TFN can produce accurate forecasts which in turn can be used in the control and operation of a system of reservoirs.

APPENDIX A17.1

ESTIMATOR FOR TFN MODELS

An estimator for obtaining MLE's for the parameters of a TFN model is outlined in this appendix. For convenience of explanation, the estimate is explained for the case of a TFN model having a single input series as in [17.2.5]. The expansion of this estimator for handling the situation where there are multiple input series as in [17.5.3], as well as interventions (see Chapter 19 and 22), is straightforward.

The estimator for a TFN model is, in fact, a direct extension of the estimator for an ARMA model. In general, the estimator works as follows. Recall from the start of Part VII that a TFN model can be qualitatively written as

single output = dynamic component + noise

where the dynamic component contains one or more input series and the correlated noise component is modelled as an an ARMA model process. As explained in Chapter 19, the dynamic component can also be designed for not only taking care of multiple input series but also the effects of external interventions upon the output series. Whatever the case, one can calculate the noise component as

noise = single output − dynamic component

Next, because the noise is assumed to be an ARMA process, one can calculate the white noise part for the correlated noise term by using an ARMA filter. Recall that these residuals are assumed to be NID$(0,\sigma_a^2)$. Finally, keeping in mind that one is simultaneously estimating the parameters for an overall TFN model, one can employ an ARMA estimator, such as the McLeod (1977) algorithm described in Appendix A6.1 or one of the other estimators listed in Section 6.2.3, to obtain MLE's for the model parameters. Possible optimization techniques for maximizing the likelihood function are also referred to in Section 6.2.3.

To be more specific, consider the TFN model in [17.2.5] having one input series, which is written as

$$\begin{aligned} y_t - \mu_y &= v(B)(x_t - \mu_x) + N_t \\ &= \frac{\omega(B)}{\delta(B)} B^b (x_t - \mu_x) + N_t \\ &= \frac{\omega(B)}{\delta(B)} (x_{t-b} - \mu_x) + N_t \end{aligned} \qquad [A17.1]$$

where y_t is the output or response series having a theoretical mean of μ_x and x_t is the input or covariate series that has a theoretical mean of μ_x. The transfer function for the input series models the influence of x_t upon y_t and is given as

$$\frac{\omega(B)}{\delta(B)}B^b = \frac{(\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_m B^m)B^b}{(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)}$$

where b is a positive integer for modelling any delay in time for x_t to affect y_t. If there is no delay effect, then $b = 0$. Lastly, because N_t is assumed to follow an ARMA process as in [3.4.4],

$$\phi(B)N_t = \theta(B)a_t$$

$$\text{or } N_t = \frac{\theta(B)}{\phi(B)}a_t \qquad [A17.2]$$

where the ARMA filter is given as

$$\frac{\theta(B)}{\phi(B)} = \frac{(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)}{(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)}$$

and the a_t's are assumed to be NID$(0,\sigma_a^2)$.

Before fitting a TFN model to the data, the response and covariate series may undergo a transformation such as the Box-Cox transformation in [3.4.30]. Whatever the case, the theoretical means of the y_t and x_t series can be estimated by the sample means given as

$$\hat{\mu}_y = \bar{y}_t = \frac{1}{n}\sum_{t=1}^{n} y_t \qquad [A17.3]$$

and

$$\hat{\mu}_x = \bar{x}_t = \frac{1}{n}\sum_{t=1}^{n} x_t \qquad [A17.4]$$

respectively, where n is the number of observations in the x_t and y_t series. One could also simultaneously estimate μ_x and μ_y along with the other model parameters, but for the length of series that are usually analyzed, the sample means provide adequate estimates. The remaining parameters to be estimated in the transfer function in the dynamic component are

$$\delta = (\delta_1, \delta_2, \ldots, \delta_r)$$

$$\omega = (\omega_0, \omega_1, \omega_2, \ldots, \omega_m)$$

where ω_0 is always included as a parameter in any transfer function. The parameters which must be estimated in the noise term are

$$\phi = (\phi_1, \phi_2, \ldots, \phi_p)$$

$$\theta = (\theta_1, \theta_2, \ldots, \theta_q)$$

Because the a_t innovations are automatically estimated during the estimation procedure, in the final iteration the variance, σ_a^2, of the innovations can be calculated as

$$\hat{\sigma}_a^2 = \frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2 \qquad [A17.5]$$

where $\hat{a}_t$ is the estimate for a_t.

Based upon the normality assumption for the innovations, one can obtain the likelihood function for a model and employ an optimization procedure for iteratively finding the values of the parameter which converge to values that maximize the likelihood function. For a given maximum likelihood estimator, at each iteration one must be able to calculate the a_t's based upon the current values of the parameters as well as the structure of the model. As explained by Box and Jenkins (1976, p. 389), one can determine the current values of the a_t's for the TFN model in [A17.1], by following a three stage procedure. Firstly, the output, d_t, from the dynamic component can be computed from

$$d_t = \frac{\omega(B)}{\delta(B)}B^b(x_t - \bar{x}) = \frac{\omega(B)}{\delta(B)}(x_{t-b} - \bar{x})$$

or

$$\delta(B)d_t = \omega(B)(x_{t-b} - \bar{x})$$

or

$$d_t - \delta_1 d_{t-1} - \cdots - \delta_r d_{t-r} = \omega_0(x_{t-b} - \bar{x}) - \omega_1(x_{t-b-1} - \bar{x})$$
$$- \cdots - \omega_m x_{t-b-m} \qquad [A17.6]$$

Next the noise, N_t, can be determined using

$$N_t = (y_t - \bar{y}) - d_t \qquad [A17.7]$$

Thirdly, the a_t's can be obtained from [A17.2] as

$$a_t = \frac{\phi(B)}{\theta(B)}N_t$$

or

$$a_t = \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} + N_t - \phi_1 N_{t-1} - \phi_2 N_{t-2} - \cdots - \phi_p N_{t-p} \qquad [A17.8]$$

In order to calculate the innovations at each iteration in the overall optimization procedures, appropriate starting values are required for x_t's, y_t's and a_t's in [A17.6] to [A17.8], respectively. As noted by Box and Jenkins (1976, p. 389), the effects of transients can be minimized if the difference equations are initiated using a value of t for which all previous x_t's and y_t's are known. Consequently d_t in [A17.6] is computed from $t = u + 1$ onwards, where u is the larger of r and $m + b$ and d_t's occurring before $u + 1$ are set equal to zero. In turn, N_t in [A17.7] can be determined from N_{u+1} onwards. Finally, if the unknown a_t's are set equal to their unconditional expected values of zero, the a_t's can be calculated from a_{u+p+1} onwards.

Since one knows how to calculate the a_t's in the ARMA noise component, an estimator developed for an ARMA model can be employed within the TFN structure to obtain MLE's for the parameter of both the noise and dynamic components. For example, one could employ the estimator of McLeod described in Appendix A6.1. The standard errors of the estimated parameters δ, ω, ϕ and θ are obtained by inverting the observed information which is obtained by numerical diferentiation of the log likelihood function.

Brockwell and Davis (1987) give an alternative method for estimating TFN models using the Kalman filter. This technique yields exact maximum likelihood estimates as opposed to the method outlined above which is approximate. In practice, however, there is normally little difference between estimates produced by the two approaches, particularly when the lengths of the series exceed 50 data points.

PROBLEMS

17.1 Select two yearly series called x_t and y_t for which you think that x_t causes y_t and then carry out the following tasks:

(a) Following the procedure of Section 16.2.2, employ the residual CCF to ascertain if your suspected causality relationship between x_t and y_t is true.

(b) Employing the three identification approaches of Section 17.3.1, design a TFN model for formally connecting x_t and y_t. Which identification method provides the most clear results and is easiest to apply?

(c) After estimating the model parameters for the TFN identified in part (b), execute suitable diagnostic checks from Section 17.3.3 and make any necessary changes to the model. Be sure to explain all of your results and write down the difference equation for the final model.

17.2 By referring to the paper of Haugh and Box (1977), outline how these authors derive their identification procedure for a TFN model.

17.3 Explain the main steps that Box and Jenkins (1976) follow to derive their TFN identification method.

17.4 Based upon your experiences in fitting TFN models to data as well as theoretical attributes of the identification procedures, compare the three identification methods described in Section 17.3.1 according to advantages and limitations.

17.5 By referring to the literature in a field of interest to you, locate three articles describing applications of TFN models. Briefly outline the types of TFN applications carried out in the papers and chapters and how the TFN modelling was of assistance to the authors.

17.6 Follow the instructions of problem 17.1 for the situation where you employ average monthly or other types of seasonal data.

17.7 Construct a TFN model for formally modelling a yearly output series for which you have two meaningful annual covariate series. For instance, you may have an average annual riverflow series as well as yearly precipitation and temperature records.

17.8 Carry out the instructions of problem 17.7 for the case of monthly or quarter monthly time series.

17.9 Obtain an average monthly riverflow data set for which you have at least two precipitation and two temperature series. Follow the approach of Section 17.5.4, to systematically select the most appropriate TFN model to link your data sets.

17.10 After reading the paper of Haltiner and Salas in which they employ the ARMAX model of Section 17.6 for short-term forecasting of snowmelt runoff, do the following:

(a) Outline the approach that they employ for modelling how streamflow is affected by other hydrological variables.

(b) Compare the procedure of Haltiner and Salas for modelling runoff to that given in Section 17.5.4.

(c) Explain how Haltiner and Salas could employ a TFN model instead of an ARMAX model to formally model their hydrological data sets.

17.11 Capodaglio et al. (1992) demonstrate that ARMAX or TFN models perform as well as deterministic differential equations for describing certain wastewater treatment processes. For their applications, explain how they accomplished this. Find and explain another physical systems problem where TFN models fare as well or better than their deterministic counterparts.

17.12 The TFN model is an example of a finite difference equation that mathematically models the relationships among data sets available at discrete time points. In continuous time, one employs stochastic differential equations. Explain the continuous-time versions of the TFN models in [17.2.5] and [17.5.3].

17.13 As pointed out in Section 17.1, Delleur (1986) demonstrates that a TFN is physically justified for modelling flows in a watershed. By referring to Delleur's paper and using both differential and difference equations, explain how he does this.

17.14 Find a paper in a field which is of interest to you that clearly explains the relevance of TFN modelling for describing physical and/or socio-economic systems. Summarize the main findings of the paper and be sure to emphasize which results you think are most interesting.

17.15 By referring to papers such as those by Novotny and Zheng (1989) and Capodaglio et al. (1990), explain how TFN models can be employed for approximately modelling nonlinear relationships between variables.

REFERENCES

For other references related to TFN modelling, the reader may wish to refer to Chapters 16, 18, 19, 20 and 22.

ARMAX MODELLING

Astrom, K. J. and Bohlin, T. (1965). Numerical identification of linear dynamic systems from normal operating records. *Proceedings of the IFAC (International Federation of Automatic Control) Symposium on Self-Adaptive Systems,* Teddington, England; also in P. H. Hammond, Editor, *Theory of Self-Adaptive Control Systems*, Plenum, New York.

Capodaglio, A. G., Jones, H., Novotny, V. and Feng, X. (1991). Sludge bulking analysis and forecasting: Application of system identification and artificial neural computing technologies. *Water Research,* 25(10):1217-1224.

Capodaglio, A. G., Novotny, V. and Fortina, L. (1992). Modelling wastewater treatment plants through time series analysis. *Environmetrics* 3(1):99-120.

Capodaglio, A. G., Zheng, S., Novotny, V. and Feng, X. (1990). Stochastic system identification of sewer flow models. *Journal of Environmental Engineering,* American Society of Civil Engineering, 116(2):284-298.

Haltiner, J. P. and Salas, J. D. (1988). Short-term forecasting of snowmelt runoff using ARMAX models. *Water Resources Bulletin,* 24(5):1083-1089.

Hannan, E. J. (1970). *Multiple Time Series.* John Wiley, New York.

Novotny, V., Jones, H., Feng, X. and Capodaglio, A. G. (1991). Time series analysis models of activated sludge plants. *Water Science Technology* 23:1107-1116.

Young, P. C. (1984). *Recursive Estimation and Time-Series Analysis.* Springer-Verlag, Berlin.

DATA SETS

Environment Canada (1979a). Historical streamflow summary, Saskatchewan to 1978. Technical report, Water Survey of Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Ontario, Canada.

Environment Canada (1979b). Historical streamflow summary, Alberta to 1978. Technical report, Water Survey of Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Ontario, Canada.

Environment Canada (1980a). Historical streamflow summary, Ontario to 1979. Technical report, Water Survey of Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Ontario, Canada.

Environment Canada (1980b). Monthly meteorological summary, Ontario to 1979. Technical report, Meteorological Branch, Environment Canada, Ottawa, Canada.

TIME SERIES ANALYSIS

Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*. John Wiley, New York.

Bartlett, M. S. (1935). *Stochastic Processes*. Cambridge University Press, London.

Beguin, J. M., Gourieroux, C. and Monfort, A. (1980). Identification of a mixed autoregressive-moving average process: the Corner method. In Anderson, O. D., Editor, *Time Series*, pages 423-436, Amsterdam. North Holland.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer-Verlag, New York.

Haugh, L. D. and Box, G. E. P. (1977). Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association*, 72(357):121-130.

Liu, L. and Hanssens, D. M. (1982). Identification of multiple-input transfer function models. *Communication in Statistics, Theory and Methods*, 11(3):297-314.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. *Biometrika*, 64(3):531-534.

Priestley, M. B. (1971). Fitting relationships between time series. *Bulletin of the International Statistical Institute*, 34:295-324.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

Vandaele, W. (1983). *Applied Time Series and Box-Jenkins Models*. Academic Press, New York.

TFN APPLICATIONS

Anselmo, V. and Ubertini, L. (1979). Transfer function-noise model applied to flow forecasting. *Hydrological Science Bulletin* IAHS (International Association of Hydrological Sciences) 24(3):353-359.

Baracos, P. C., Hipel, K. W. and McLeod, A. I. (1981). Modelling hydrologic time series from the Arctic. *Water Resources Bulletin*, 17(3):414-422.

Campbell, A., Noakes, D. J. and Elner, R. W. (1991). Temperature and lobster, homarus americanus, yield relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 48(11):2073-2082.

Chow, K. C. A., Watt, W. E. and Watts, D. G. (1983). A stochastic-dynamic model for real time flood forecasting. *Water Resources Research*, 19(3):746-752.

Delleur, J. W. (1986). Recursive parameter identification for flash flood forecasting. In *Multivariate Analysis of Hydrologic Processes*, Proceedings of the Fourth International Hydrology Symposium held at Colorado State University, Fort Collins, Colorado, July 15 to 17, 1985. Edited by H. W. Shen, J. T. B. Obeysekera, V. Yevjevich and D. G. DeCoursey, published by the Engineering Research Center, Colorado State University, 154-177.

Fay, D. M., Watt, W. E. and Watts, D. G. (1987). A stochastic real-time spring flood forecasting system for Carman, Manitoba. *Canadian Journal of Civil Engineering*, 14(1):87-96.

Gurnell, A. M. and Fenn, C. R. (1984). Box-Jenkins transfer function models applied to suspended sediment concentration-discharge relationships in a proglacial stream. *Arctic and Alpine Research* 16(1):93-106.

Hipel, K. W., McLeod, A. I. and Li, W. K. (1985). Causal and dynamic relationships between natural phenomena. In Anderson, O. D., Ord, J. K. and Robinson, E. A., Editors, *Time Series Analysis: Theory and Practice*, pages 13-34. North-Holland, Amsterdam.

Hipel, K. W., McLeod, A. I. and McBean, E. A. (1977). Stochastic modelling of the effects of reservoir operation. *Journal of Hydrology*, 32:97-113.

Hipel, K. W., McLeod, A. I. and Noakes, D. J. (1982). Fitting dynamic models to hydrological time series. In El-Shaarawi, A. H. and Esterby, S. R., Editors, *Time Series Methods in Hydrosciences*, 110-129. Elsevier, Amsterdam.

Lemke, K. A. (1990). An evaluation of transfer function/noise models of suspended sediment concentration. *The Professional Geographer* 42(3):324-335.

Lemke, K. A. (1991). Transfer function models of suspended sediment concentration. *Water Resources Research*, 27(3):293-305.

Li, W. K. (1981). *Topics in Time Series Modelling*. PhD thesis, Dept. of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada.

Maidment, D. R., Miaou, S. P. and Crawford, M. M. (1985). Transfer function models of daily urban water use. *Water Resources Research*, 21(4):425-432.

Noakes, D. J., Welch, D. W. and Stocker, M. (1987). A time series approach to stock-recruitment analysis: transfer function noise modelling. *Natural Resource Modeling*, 2(2):213-233.

Novotny, V. and Zheng, S. (1989). Rainfall-runoff function by ARMA modeling. *Journal of Hydraulic Engineering*, ASCE (American Society of Civil Engineering) 115(10):1386-1400.

Olason, T. and Watt, W. E. (1986). Multivariate transfer function-noise model of river flow for hydropower operation. *Nordic Hydrology*, 17:185-202.

Snorrason, A., Newbold, P. and Maxwell, W. H. C. (1984). Multiple input transfer function-noise modeling of river flow. In Maxwell, W. H. C. and Beard, L. R., Editors, *Frontiers in Hydrology*, pages 111-126. Water Resources Publications, Littleton, Colorado.

Thompstone, R. M. (1983). *Topics in Hydrological Time Series Modelling*. PhD thesis, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1983). Transfer function-noise modelling for power-house inflow forecasting. *INFOR*, 21(4):258-269.

Viessman Jr., W., Knapp, J. W., Lewis, G. L. and Harbaugh, T. E. (1977). *Introduction to Hydrology*. Harper and Row, New York, second edition.

Welch, D. W. and Noakes, D. J. (1991). Optimal harvest rate policies for rebuilding the Adams River sockeye salmon (oncorhynchus nerka). *Canadian Journal of Fisheries and Aquatic Sciences*, 48(4):526-535.

CHAPTER 18

FORECASTING WITH TRANSFER FUNCTION-NOISE MODELS

18.1 INTRODUCTION

A *transfer function-noise (TFN)* model can describe the dynamic relationship between a single output series and one or more input series. For example, a TFN model can formally specify the mathematical association existing between riverflows and the temperature and precipitation variables which caused the flows. Furthermore, the remaining noise component can be modelled using an ARMA model. Because its inherent flexible design reflects many types of physical situations that can take place in practice, the TFN model constitutes an important tool for use in water resources and environmental engineering plus many other fields.

In the previous chapter, the TFN model is defined and comprehensive model construction techniques are presented so that the model can be conveniently applied in practice. Moreover, practical applications are given in Chapter 17 to explain clearly how model building is carried out. If one is confronted with a situation where the direction of causality between two series is not clear, the residual cross-correlation function (CCF) of Section 16.2 can be utilized. Additionally, as explained in Section 17.3.1, after the type of causality is established, the results of a residual cross-correlation function study can be employed for deciding upon the parameters to include in a formal mathematical model to describe the relationship between the two series.

A particularly useful and common application of a calibrated TFN noise model is *forecasting*. For instance, forecasts of riverflows based upon other previous flows as well as other hydrological conditions are useful for optimizing the operation of multipurpose reservoir systems. Consequently, the objective of this chapter is to demonstrate the utility of TFN models in forecasting by employing practical applications in hydrology.

In the next section, it is explained how *minimum mean square error (MMSE) forecasts* can be generated using a TFN model. Then, *practical forecasting applications* are presented in the subsequent two sections. The forecasting experiments of Section 18.3 demonstrate that TFN models produce more accurate forecasts than other competing models, including what is called a conceptual hydrological model. The forecasting applications of Section 18.4 explain how forecasts from TFN and other models can be combined in an optimal fashion in an attempt to obtain improved forecasts. In particular, a TFN, periodic autoregressive (PAR) (see Chapter 14) and a conceptual model (see Section 18.3.3) are employed to forecast quarter monthly riverflows. These models all approach the modelling and forecasting problem from three different perspectives and each has its own particular strengths and weaknesses. The forecasts generated by the individual models are combined in an effort to exploit the strengths of each model. The results of this case study indicate that significantly better forecasts can be obtained when forecasts from different types of models are combined. The forecasting findings of Sections 18.3 and 18.4 are based upon research by Thompstone et al. (1985) and McLeod et al. (1987), respectively.

Because TFN models have been found to produce reliable forecasts in applications, they are becoming popular with practitioners. In addition to the forecasting studies described in this book, other documented results of TFN forecasting include contributions in hydrology (Anselmo and Ubertini, 1979; Baracos et al. 1981; Chow et al., 1983; Snorrason et al., 1984; Alley, 1985; Maidment et al., 1985; Olason and Watt, 1986; Fay et al., 1987; Haltiner and Salas, 1988), fish population studies (Stocker and Noakes, 1988; Noakes et al., 1990; Schweigert and Noakes, 1990) as well as many other fields. Moreover, as explained in Section 18.5, TFN models can also be employed for extending time series records, control and simulation.

For forecasting with nonseasonal ARMA and ARIMA models, the reader may wish to refer to Chapter 8. Forecasting experiments are presented in Section 15.5 on the three types of seasonal models from Part VI. These seasonal forecasting studies include experiments on combining forecasts from different seasonal models to try to procure better forecasts.

18.2 FORECASTING PROCEDURES FOR TFN MODELS

18.2.1 Overview

A TFN model describes mathematically how one or more inputs dynamically affect a single output or response variable. In Section 17.2, a TFN model having one input or covariate series is defined in [17.2.5]. Within Section 17.5.2, a TFN model with two or more covariate series is given in [17.5.3].

Intuitively, one would expect that forecasts for the response series should be considerably improved if one uses forecasting information coming from the covariate series. Consequently, the forecasts from a TFN model should be more accurate than those obtained from a separate time series model fitted only to the response series. In fact, the forecasting experiments of Section 18.3 demonstrate that a TFN model forecasts better than other competing time series models as well as a conceptual model. When a response variable can be anticipated by changes in the values of a covariate, economists refer to the covariate as a *leading indicator* for the response. The future net growth in a national economy, for instance, is often anticipated by leading indicators such as trade surplus or deficits, interest rates, unemployment and inflation.

Section 8.2 explains how to calculate minimum mean square error (MMSE) forecasts for nonseasonal ARMA and ARIMA models, while Section 15.2 describes how to compute MMSE forecasts for three types of seasonal models. The purpose of this section is to present procedures for determining MMSE forecasts for various types of TFN models. More specifically, in Section 18.2.2, formulae are given for calculating MMSE forecasts for TFN models having single or multiple inputs, ARMA or ARIMA noise and a deterministic trend component. Moreover, these kinds of TFN models can be fitted to yearly or deseasonalized data sets that may first be transformed using a Box-Cox transformation. In Section 18.2.3, an illustrative forecasting application is presented for clearly explaining how to calculate MMSE forecasts and for demonstrating that a TFN model forecasts more accurately than an ARMA model separately fitted to the response series.

18.2.2 Forecasting Formulae

For convenience of explanation, forecasting formulae are first developed for the case of a TFN model having a single covariate series. As explained below, these formulae can easily be extended for handling situations for which there are two or more input series. Other complications that are discussed in this subsection include how to handle seasonality, differencing and trends when forecasting with a TFN model.

Single Input TFN Model Having ARMA Noise

Derivation of MMSE Forecasts: As in Section 17.2, suppose that a variable X causes a variable Y. Let the observations for X and Y at time t be given by X_t and Y_t, respectively. If the given series are transformed using a transformation such as the Box-Cox transformation in [3.4.30], let the transformed series for X_t and Y_t be denoted as x_t and y_t, respectively. As in [17.2.5], a TFN model for mathematically describing the relationship between x_t and y_t as well as the noise, is written as

$$y_t - \mu_y = v(B)(x_t - \mu_x) + N_t \qquad [18.2.1]$$

where μ_y and μ_x are the theoretical means of y_t and x_t, respectively. In the above equation,

$$v(B) = \frac{\omega(B)}{\delta(B)} = \frac{(\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_m B^m)}{(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)} \qquad [18.2.2]$$

is the transfer function which models the dynamic effects of the input upon the output. If there is a delay time, b, (where b is a positive integer) for x_t to affect y_t, then x_t is replaced by x_{t-b} in [18.2.1]. The noise term, N_t, is assumed to follow an ARMA process as in [3.4.4] such that

$$\phi(B)N_t = \theta(B)a_t$$

or

$$N_t = \frac{\theta(B)}{\phi(B)}a_t = \frac{(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)}{(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)}a_t \qquad [18.2.3]$$

As pointed out later, N_t could also be an ARIMA model when the data are nonstationary.

As in [16.2.3], suppose that x_t can be described using an ARMA model such that

$$\phi_x(B)(x_t - \mu_x) = \theta_x(B)u_t$$

or

$$(x_t - \mu_x) = \frac{\theta_x(B)}{\phi_x(B)}u_t$$

$$= \frac{(1 - \theta_{x,1}B - \theta_{x,2}B^2 - \cdots - \theta_{x,q_x}B^{q_x})}{(1 - \phi_{x,1}B - \phi_{x,2}B^2 - \cdots - \phi_{x,p_x}B^{p_x})}u_t \qquad [18.2.4]$$

By substituting the above into [18.2.1], the TFN model becomes

$$y_t - \mu_y = \frac{\omega(B)\theta_x(B)}{\delta(B)\phi_x(B)}u_t + \frac{\theta(B)}{\phi_{(B)}}a_t$$

$$= v^*(B)u_t + \psi(B)a_t \qquad [18.2.5]$$

The transfer function for u_t in [18.2.5] is expanded as

$$v^*(B) = v_0^* + v_1^*B + v_2^*B^2 + \cdots$$

where the v^* weights can be calculated by equating coefficients in the identity

$$v^*(B) = \frac{\omega(B)\theta_x(B)}{\delta(B)\phi_x(B)}$$

or

$$\delta(B)\phi_x(B)v^*(B) = \omega(B)\theta_x(B) \qquad [18.2.6]$$

As in [3.4.18], the random shock operator is

$$\psi(B) = \frac{\theta(B)}{\phi(B)} = 1 + \psi_1B + \psi_2B^2 + \cdots \qquad [18.2.7]$$

where the ψ_i weights can be determined using the identity in [3.4.21].

By replacing t by $t+l$ in [18.2.5], the TFN model for the actual value of the response variable at time $t+l$ is

$$\begin{aligned} y_{t+l} - \mu_y = &\Big(v_0^*u_{t+l} + v_1^*u_{t+l-1} + v_2^*u_{t+l-2} + \cdots \\ &+ v_l^*u_t + v_{l+1}^*\mu_{t-1} + v_{l+2}^*u_{t-2} + \cdots\Big) \\ &+ \Big(a_{t+l} + \psi_1a_{t+l-1} + \psi_2a_{t+l-2} + \cdots \\ &+ \psi_la_t + \psi_{l+1}a_{t-1} + \psi_{l+2}a_{t-2} + \cdots\Big) \end{aligned} \qquad [18.2.8]$$

where l is a positive integer. Let $\hat{y}_t(l)$ be the forecast for y_{t+l} made at origin t. Keeping in mind that only information up to time t can be utilized, let this forecast be written as

$$\begin{aligned} \hat{y}_t(l) - \mu_y = &(v_l^o\mu_t + v_{l+1}^ou_{t-1} + v_{l+2}^ou_{t-2} + \cdots) \\ &+ (\psi_l^oa_t + \psi_{l+1}^oa_{t-1} + \psi_{l+2}^oa_{t-2} + \cdots) \end{aligned} \qquad [18.2.9]$$

Then, using [18.2.8] and [18.2.9]

$$\begin{aligned} y_{t+l} - \hat{y}_t(l) = &\sum_{i=0}^{l-1}(v_i^*u_{t+l-i} + \psi_ia_{t+l-i}) \\ &+ \sum_{j=0}^{\infty}\left[(v_{l+j}^* - v_{l+j}^o)u_{t-j} + (\psi_{l+j} - \psi_{l+j}^o)a_{t-j}\right] \end{aligned} \qquad [18.2.10]$$

where $\psi_0 = 1$. Following arguments put forward in Section 8.2.2 for forecasting with an ARMA model, one can determine the MMSE forecast for the response variable. In particular, the mean square error for the forecast is calculated using [18.2.10] within the expected value given below as

$$\begin{aligned}E[y_{t+l}-\hat{y}_t(l)]^2 &= (\nu_0^{*2}+\nu_1^{*2}+\nu_2^{*2}+\cdots+\nu_{l-1}^{*2})\sigma_u^2\\ &+(1+\psi_1^2+\psi_2^2+\cdots+\psi_{l-1}^2)\sigma_a^2\\ &+\sum_{j=0}^{\infty}[(\nu_{l+j}^{*}-\nu_{l+j}^{o})^2\sigma_u^2+(\psi_{l+j}-\psi_{l+j}^{o})^2\sigma_a^2]\end{aligned} \qquad [18.2.11]$$

which is minimized only if $\nu_{l+j}^o=\nu_{l+j}^*$ and $\psi_{l+j}^o=\psi_{l+j}$. Consequently, the MMSE forecast $\hat{y}_t(l)$ of y_{t+l} at origin t is given by the *conditional expectation* of y_{t+l} at time t. Therefore, the MMSE forecast using the TFN model as written in [18.2.9] is simply

$$\begin{aligned}\hat{y}_t(l)-\mu_y &= (\nu_l^* u_t+\nu_{l+1}^* u_{t-1}+\nu_{l+2}^* u_{t-2}+\cdots)\\ &+(\psi_l a_t+\psi_{l+1}a_{t-1}+\psi_{l+2}a_{t-2}+\cdots)\end{aligned} \qquad [18.2.12]$$

Computing MMSE Forecasts: Equation [18.2.12] could be employed for calculating MMSE forecasts for a TFN model having a single input. However, a more convenient way to compute the forecasts is to use the TFN format from [18.2.1] which is written at time $t+l$ as

$$y_{t+l}-\mu_y=\frac{\omega(B)}{\delta(B)}(x_{t+l}-\mu_x)+\frac{\theta(B)}{\phi(B)}a_{t+l} \qquad [18.2.13]$$

when the noise is modelled as an ARMA process. To eliminate the operators written in the denominators on the right hand side of the equation, one can multiply both sides of the equation by $\phi(B)\delta(B)$ to obtain

$$\phi(B)\delta(B)(y_{t+l}-\mu_y)=\phi(B)\omega(B)(x_{t+l}-\mu_x)+\delta(B)\theta(B)a_{t+l} \qquad [18.2.14]$$

Subsequently, one can multiply together the operators in each term in [18.2.13] and then take conditional expectations to determine the MMSE forecasts. Specifically, in [18.2.13] let

$$\begin{aligned}\delta^*(B)&=\phi(B)\delta(B)=1-\delta_1^*B-\delta_2^*-\cdots-\delta_{p+r}^*B^{p+r}\\ \omega^*(B)&=\phi(B)\omega(B)=1-\omega_0^*-\omega_1^*-\cdots-\omega_{p+s}^*B^{p+s}\\ \theta^*(B)&=\delta(B)\theta(B)=1-\theta_1^*B-\theta_2^*B^2-\cdots-\theta_{q+r}^*B^{q+r}\end{aligned} \qquad [18.2.15]$$

One can see that δ_i^*, ω_i^* and θ_i^* coefficients can be easily computed by multiplying together the known operators as defined above. Then, employing square brackets to denote conditional expectations at time t, the MMSE forecast for lead time l is

$$\begin{aligned}\hat{y}_t(l)-\mu_y &= [y_{t+l}]-\mu_y\\ &=\delta_1^*([y_{t+l-1}]-\mu_y)+\delta_2^*([y_{t+l-2}]-\mu_y)\\ &+\cdots+\delta_{p+r}^*([y_{t+l-p-r}]-\mu_y)\end{aligned}$$

$$+ \omega_o^*([x_{t+l}] - \mu_x) - \omega_1^*([x_{t+l-1}] - \mu_x)$$
$$- \omega_2^*([x_{t+l-2}] - \mu_x) - \cdots - \omega_{p+s}^*([x_{t+l-p-s}] - \mu_x) + [a_{t+l}]$$
$$- \theta_1^*[a_{t+l-1}] - \theta_2^*[a_{t+l-2}] - \cdots - \theta_{q+r}^*[a_{t+l-q-r}] \qquad [18.2.16]$$

In order to obtain the MMSE forecasts, the rules for iteratively calculating the conditional expectations in [18.2.5] for lead times $l = 1,2,\ldots,$ are as follows:

$$[y_{t+j}] = \begin{cases} y_{t+j} & \text{for } j \le 0 \\ \hat{y}_t(j) & \text{for } j > 0 \end{cases} \qquad [18.2.17a]$$

since y_{t+j} is a known observation for $j \le 0$ and unknown for $j > 0$.

$$[x_{t+j}] = \begin{cases} x_{t+j} & \text{for } j \le 0 \\ \hat{x}_t(j) & \text{for } j > 0 \end{cases} \qquad [18.2.17b]$$

where the forecasts for the input variable are determined using the ARMA model for the x_t series in [18.2.4] according to the forecasting rules laid out in Section 8.2.4 for an ARMA model.

$$[a_{t+j}] = \begin{cases} a_{t+j} & \text{for } j \le 0 \\ 0 & \text{for } j > 0 \end{cases} \qquad [18.2.17c]$$

because a_{t+j} is known for $j \le 0$ and has an expected value of zero for $j > 0$.

Variance of MMSE Forecasts: To obtain the v_i^* and ψ_i weights for the TFN model as written in [18.2.5], one can employ the identities in [18.2.6] and [3.4.21], respectively. On the right hand side of [18.2.10], the forecast error is given by the first summation component. From the first two terms on the right hand side of [18.2.11], the variance of the forecast error for lead time l is written as

$$V(l) = E[y_{t+l} - \hat{y}_t(l)]^2$$
$$= \sigma_\mu^2 \sum_{j=0}^{l-1} v_j^{*2} + \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 \qquad [18.2.18]$$

where σ_μ^2 is the variance of the noise term for the ARMA model fitted to the x_t series in [18.2.4], σ_a^2 is the variance of the ARMA noise term for the TFN model in [18.2.1] and [18.2.12], and the v_j^* and ψ_i coefficients are determined using the identities in [18.2.6] and [3.4.21], respectively. When the μ_t and a_t series are assumed to be NID$(0,\sigma_\mu^2)$ and NID$(0,\sigma_a^2)$, respectively, one can conveniently calculate the probability limits for each MMSE forecast. For instance, the 95% probability limits for $\hat{y}_t(l)$ would be $\hat{y}_t(l) \pm 1.96\sqrt{\hat{V}(l)}$ where $\hat{V}(l)$ is estimated using [18.2.18] when the coefficient and noise estimates appropriately replace the theoretical values given on the right hand side of the equation.

Forecasts in the Original Domain: If the y_t or x_t are transformed using a Box-Cox transformation from [3.4.30], the MMSE forecasts calculated above are for the transformed domain. To obtain forecasts in the original units or untransformed domain, one would have to take an inverse Box-Cox transformation as is explained in Section 8.2.7 for the case of an ARMA model fitted to a single series. Keep in mind that both the forecasts and corresponding probability limits in the transformed domain can be determined for the untransformed domain.

Multiple Input TFN Model Having ARMA Noise

The TFN model in [17.2.5] and [18.2.1] has a single covariate or input series x_t. In general, one could have a TFN model with I input series which is written in [17.5.3] as

$$\begin{aligned}(y_t - \mu_y) &= \frac{\omega_1(B)}{\delta_1(B)}(x_{t_1} - \mu_{x_1}) \\ &+ \frac{\omega_2(B)}{\delta_2(B)}(x_{t_2} - \mu_{x_2}) + \cdots + \frac{\omega_I(B)}{\delta_I(B)}(x_{t_I} - \mu_{x_I}) \\ &+ \frac{\theta(B)}{\phi(B)}a_t \qquad [18.2.19]\end{aligned}$$

To employ this model for forecasting purposes, one follows a procedure similar to that carried out for the TFN model having a single covariate series. In particular, first one must write the TFN model so that there is no operator in the denominator of any term in [18.2.19]. This is accomplished by multiplying [18.2.19] by $\{\delta_1(B)\delta_2(B)\cdots\delta_I(B)\phi(B)\}$. Next, by separately fitting an ARMA model to each of the x_{t_i} series, one calculates the MMSE forecasts for each x_{t_i} series by following the procedure of Section 8.2.4. Thirdly, one iteratively calculates the MMSE forecasts for the response or output series for lead times $l = 1,2,\ldots,$ using the rules in [8.2.16]. Additionally, using a formula similar to that given in [18.2.18] for a TFN having one input, one can determine the probability limits for each forecast. Finally, if the response variable and other input covariates have been transformed using a Box-Cox transformation, one can, if desired, calculate the corresponding forecasts and probability limits in the untransformed domain.

Seasonal TFN Model

As noted in Section 17.2.1, a simple procedure is available for handling seasonal data. Firstly, the output series and each of the input series may be transformed using a Box-Cox transformation in order to cause each time series to be approximately normally distributed. Secondly, assuming that there is approximate stationarity within each season for a given series so that a graph of the series follows a shape similar to that in Figure VI.1 for an average monthly riverflow series, one can deseasonalize the series using a procedure from Section 13.2.2. Next, an appropriate TFN model is fitted to the set of deseasonalized series using the model construction techniques explained in Sections 17.3 and 17.5.3, and an ARMA model is separately developed for each deseasonalized input series by following the model building methods of Part III. Fourthly, by employing the forecasting formulae of Section 8.2.4, MMSE forecasts can be iteratively generated for each deseasonalized input series using the ARMA model fitted to the series. Next, by utilizing the forecasting formulae for TFN models presented in this section as well as the forecasts for the inputs, MMSE forecasts can be iteratively determined for the response variable for lead times $l = 1,2, \ldots,$. Finally, to obtain forecasts in the untransformed domain, one must first take the inverse deseasonalization transformation of the forecasts and then invoke the inverse Box-Cox transformation. The procedure for forecasting seasonal data using a TFN model is depicted in Figure 18.2.1.

Another approach for handling seasonal data is to employ a periodic TFN model. The interested reader may wish to explore this possibility by answering problem 18.5.

TFN Model Having ARIMA Noise

Suppose that one wishes to fit a TFN model to a nonseasonal response series which has one nonseasonal input series and that these two series are nonstationary. One way to remove this nonstationarity is to introduce the differencing operator of Section 4.3.1 into the noise term of the TFN model so that the noise component follows an ARIMA rather than an ARMA process. Accordingly, the nonstationary version of the TFN model in [18.2.1] is

$$y_t = \frac{\omega(B)}{\delta(B)}x_t + \frac{\theta(B)}{\phi(B)\nabla^d}a_t \qquad [18.2.20]$$

where

$$\nabla^d = (1-B)^d$$

is the differencing operator defined for d taking on values of zero when the data are stationary and positive integers when the data are nonstationary. As exemplified by the examples in Section 4.3.3, usually $d = 1$ or 2 when an ARIMA model is fitted to a single yearly nonstationary time series. Because of the differencing operator in [18.2.20], both the y_t and x_t series are assumed not to have mean levels.

To obtain MMSE forecasts for the TFN model in [18.2.20], the procedure is similar to that for the stationary case. Firstly, one must eliminate operators in the denominator by multiplying [18.2.20] by $\delta(B)\phi(B)\nabla^d$ to obtain

$$\delta(B)\phi(B)\nabla^d y_t = \phi(B)\nabla^d\omega(B)x_t + \delta(B)\theta(B)a_t \qquad [18.2.21]$$

Next, after multiplying together the operators in each term in [18.2.21], one can iteratively calculate the MMSE forecasts by employing the rules in [18.2.17]. Finally, after taking into account the fact that there is a differencing operator, probability limits can be calculated for each forecast using a formula similar to that in [18.2.18].

If one were dealing with seasonal time series that follow graphs similar to those in Figures VI.2 or VI.3, one could possibly model a set of these time series using a TFN model by having a SARIMA noise term. As defined in Section 12.2.1, a SARIMA model contains nonseasonal and seasonal differencing operators to remove nonseasonal and seasonal nonstationarity, respectively. Moreover, the SARIMA model also has seasonal AR and MA operators in addition to the nonseasonal AR and MA operators.

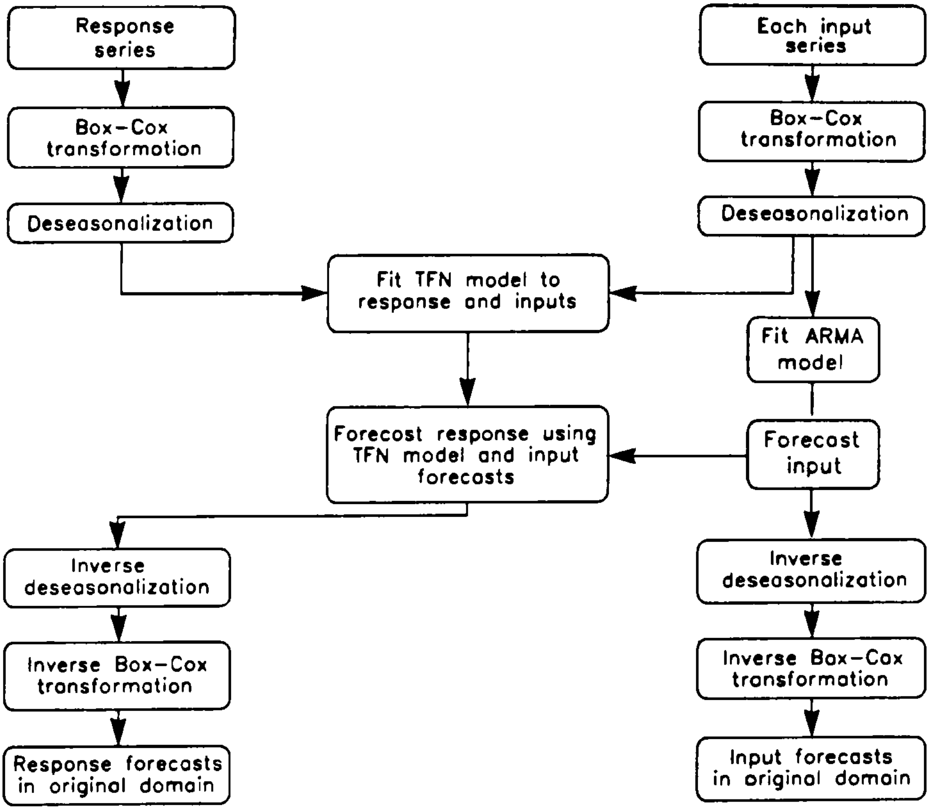

Response series
Box-Cox transformation
Deseasonalization
Each input series
Box-Cox transformation
Deseasonalization
Fit TFN model to response and inputs
Fit ARMA model
Forecast response using TFN model and input forecasts
Forecast input
Inverse deseasonalization
Inverse Box-Cox transformation
Response forecasts in original domain
Inverse deseasonalization
Inverse Box-Cox transformation
Input forecasts in original domain

Figure 18.2.1. Forecasting seasonal series using a TFN model.

TFN Model Having a Deterministic Trend

As discussed in Sections 4.5 and 4.6, differencing is designed for removing stochastic trends in a time series. However, differencing may not eliminate a deterministic trend contained in a time series. To explain how a deterministic trend can be modelled and forecasted rewrite [18.2.20] as

$$\nabla^d y_t - \mu_w = \frac{\omega(B)}{\delta(B)}\nabla^d x_t + \frac{\theta(B)}{\phi(B)}a_t \qquad [18.2.22]$$

where μ_w is a level in the y_t series that the differencing cannot eliminate. By multiplying [18.2.22] by $\delta(B)\phi(B)$, one obtains

$$\delta(B)\phi(B)\nabla^d y_t = \theta_0 + \phi(B)\omega(B)\nabla^d x_t + \delta(B)\theta(B)a_t \qquad [18.2.23]$$

where

$$\theta_0 = \delta(B)\phi(B)\mu_w$$
$$= \delta(1)\phi(1)\mu_w$$

In the expression for θ_0, one replaces each B by unity in the two operators because $B^k\mu_w = \mu_w$ for $k = 0,1,2,\cdots$.

The procedure for calculating MMSE forecasts using [18.2.23] is the same as before except for the θ_0 term on the right hand side of [18.2.23]. Consequently, the rules for conditional expectations in [18.2.17] are employed to determine the MMSE forecasts for the response and input series, keeping in mind that $[\theta_0] = \theta_0$ in [18.2.23].

18.2.3 Application

The Red Deer River is a tributary of the South Saskatchewan (abbreviated as S.Sask.) River which flows eastwards from the Rocky Mountains across the Canadian prairies. In Section 17.4.2, a TFN model is constructed for describing the influence of the deseasonalized logarithmic Red Deer riverflows upon the deseasonalized logarithmic S.Sask. riverflows. From [17.4.1], this calibrated TFN model is written as

$$y_t = (0.572 + 0.238B)x_t + \frac{(1 - 0.494B)}{(1 - 0.856B)}a_t \qquad [18.2.24]$$

where y_t and x_t are the deseasonalized logarithmic S.Sask. and Red Deer riverflows, respectively, and $\hat{\sigma}_a^2 = 0.310$.

To write the model in [18.2.24] in a convenient form to calculate MMSE forecasts, first multiply the difference equation by the operator $(1 - 0.856B)$ to get

$$(1 - 0.856B)y_t = (1 - 0.856B)(0.572 + 0.238B)x_t + (1 - 0.494B)a_t$$

or

$$y_t = 0.856y_{t-1} + 0.572x_t - 0.252x_{t-1} - 0.204x_{t-2} + a_t - 0.494a_{t-1}$$

By replacing t by t+l and taking conditional expectations in the above equations, the MMSE forecast for lead time l is

$$\hat{y}_t(l) = 0.856[y_{t+l-1}] + 0.572\hat{x}_t(l) - 0.252[x_{t+l-1}] - 0.204[x_{t+l-2}]$$
$$+ [a_{t+l}] - 0.494[a_{t+l-1}] \qquad [18.2.25]$$

For the case of a one-step-ahead forecast where $l = 1$ the above equation becomes

$$\hat{y}_t(1) = 0.856[y_t] + 0.572\hat{x}_t(1) - 0.252[x_t] - 0.204[x_{t-1}] + [a_{t+1}] - 0.494[a_t]$$
$$= 0.856y_t + 0.572\hat{x}_t(1) - 0.252x_t - 0.204x_{t-1} - 0.494a_t \qquad [18.2.26]$$

Notice in [18.2.25] and [18.2.26], one requires MMSE forecasts for the x_t series. Consequently, one must separately fit an ARMA model to the x_t series and then use this model to generate MMSE forecasts for the x_t series. When an ARMA model is separately designed for describing the x_t series for the deseasonalized logarithmic Red Deer flows, the most appropriate ARMA model is found to be an ARMA(1,1). The estimated ARMA(1,1) model for the x_t series is given as

$$(1 - 0.845B)x_t = (1 - 0.292B)u_t \qquad [18.2.27]$$

where u_t is the innovation series at time t and $\hat{\sigma}_u^2 = 0.482$. By substituting $t+l$ for t and taking conditional expectations in [18.2.27], the formula for iteratively generating MMSE forecasts for x_t is

$$\hat{x}_t(l) = 0.845[x_{t+l-1}] + [u_{t+l}] - 0.292[u_{t+l-1}] \qquad [18.2.28]$$

To obtain the one-step-ahead MMSE forecast in [18.2.28] simply replace l by unity to obtain

$$\hat{x}_t(1) = 0.845[x_t] + [u_{t+1}] - 0.292[u_t]$$
$$= 0.845x_t - 0.292u_t \qquad [18.2.29]$$

To calculate MMSE forecasts for the y_t series in the TFN model in [18.2.25], one can employ [18.2.28] to determine the MMSE forecasts for the x_t series which are needed as input forecasts in [18.2.25]. Consider the case where one wishes to find $\hat{y}_t(1)$ using [18.2.26]. Firstly, $\hat{x}_t(1)$ is found by utilizing [18.2.29] and then $\hat{x}_t(1)$ is substituted into [18.2.26] to get $\hat{y}_t(1)$. From [18.2.18], the variance of the one-step-ahead MMSE forecast error for $\hat{y}_t(1)$ is given theoretically as

$$V(1) = \omega_0^2\sigma_u^2 + \sigma_a^2 \qquad [18.2.30]$$

where the estimate is calculated as

$$\hat{V}(1) = (0.572)^2 0.482 + 0.310 = 0.468$$

To calculate the lead one MMSE forecast for the original untransformed series, the deseasonalization and logarithmic transformations must be taken into account. Accordingly, from time t, the lead one MMSE forecast $\hat{y}_t(1)$ for the untransformed series is estimated using

$$\hat{y}_t(1) = \exp\left(\hat{y}_t(1)\hat{\sigma}_m + \hat{\mu}_m + \frac{1}{2}(0.468)\hat{\sigma}_m^2\right) \qquad [18.2.31]$$

where $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the estimated mean and standard deviation calculated using [13.2.4] and [13.2.5], respectively, for the month that is currently connected with time t+1 of the y_t series. As explained in Section 8.2.7, the last term in the exponent in [18.2.31] is the correction required for producing the MMSE forecast in the untransformed domain.

When calculating MMSE forecasts for the y_t series using the TFN model in [18.2.25], information from the input series is used in the forecast calculation. Consequently, a priori, one would expect a TFN model to forecast more accurately than an ARMA model that is separately fitted to the response series. For the case of the y_t series representing the deseasonalized logarithmic flows of the S. Sask. River, the most appropriate model to fit to this series is an ARMA(1,1) model which is calibrated as

$$(1 - 0.819)y_t = (1 - 0.253)v_t \qquad [18.2.32]$$

where v_t is the innovation series at time t and $\hat{\sigma}_v^2 = 0.507$. Notice that the variance of the noise has a value of $\hat{\sigma}_v^2 = 0.507$ for the ARMA model in [18.2.32] and a magnitude of $\hat{\sigma}_a^2 = 0.310$ for the TFN model in [18.2.24]. Consequently, the TFN model provides a better fit to the available information than the single ARMA model and has a residual variance which is about 40% smaller. By replacing t by t+l and taking conditional expectations in [18.2.32], the formula for iteratively determining MMSE forecasts for y_t using an ARMA model is

$$\hat{y}_t(l) = 0.819[y_{t+l-1}] + [v_{t+l}] - 0.253[v_{t+l-1}] \qquad [18.2.33]$$

To ascertain the one-step-ahead MMSE forecast in [18.2.33], simply assign l a value of one to get

$$\begin{aligned}\hat{y}_t(1) &= 0.819[y_t] + [v_{t+l}] - 0.253[v_t]\\ &= 0.819y_t - 0.253v_t \qquad [18.2.34]\end{aligned}$$

In the untransformed domain, the lead one MMSE forecast, $\hat{y}_t(1)$ is calculated using

$$\hat{y}_t(1) = \exp\left(\hat{y}_t(1)\hat{\sigma}_m + \hat{\mu}_m + \frac{1}{2}(0.507)\sigma_m^2\right) \qquad [18.2.35]$$

where $\hat{y}_t(1)$ is determined using [18.2.34], $\hat{\sigma}_v^2 = 0.507$ as in [18.2.32], and $\hat{\sigma}_m$ as well as $\hat{\mu}_m$ are the same as in [18.2.31].

As would be expected the TFN model for the response variable produces more accurate forecasts than an ARMA model separately fitted to the same series. More specifically, when the TFN model in [18.2.26] along with the inverse transformation in [18.2.31] are employed for obtaining lead one MMSE forecasts in the untransformed domain, the mean square error for all months in 1963 is about 20% less than for the forecasts obtained using the ARMA(1,1) model in [18.2.34] and [18.2.35] for the output series. Consequently, the input series in the TFN model acts as a leading indicator to significantly improve the accuracy of the MMSE forecasts of the output series.

18.3 FORECASTING QUARTER-MONTHLY RIVERFLOWS

18.3.1 Overview

A key problem in the operation of a water resources system is the forecasting of natural inflows to the various reservoirs in the system. It is increasingly recognized that time series analysis is of considerable practical use in dealing with this problem. The current section demonstrates the practical importance of this methodology by examining the use of the TFN models of Chapter 17 to forecast natural inflows in the Lac St. Jean reservoir, a major component of the multi-reservoir hydroelectric system operated by Alcan Smelters and Chemicals Ltd. in the province of Quebec, Canada.

The electricity generated by this system is used at Alcan's aluminum smelter in Arvida, Quebec. In order to insure a constant and adequate supply of power, it is necessary to schedule releases from the reservoir in an optimum fashion. Thus, forecasts of the quarter monthly inflows into the reservoir are required so that the desired outflow and hydraulic head are available for power generation.

The forecasting experiments presented in this section were originally presented by Thompstone et al. (1985). The output for the TFN model used in the study are the quarter-monthly (i.e., near-weekly) natural inflows to the Lac St. Jean reservoir. The covariate series for the TFN model are rainfall and snowmelt, the latter being a novel derivation from daily rainfall, snowfall and temperature series. It is clearly demonstrated in Section 18.3.2 using the residual variance and the AIC (see Section 6.3) that modelling is improved as one starts with a deseasonalized ARMA model (Chapter 13) of the inflow series and successively adds transfer functions for the rainfall and snowmelt series. It is further demonstrated that the TFN model is better than a PAR model (Chapter 14) of the inflow series. The split-sample experiments are used in Section 18.3.4 to compare one-step-ahead forecasts from this TFN model with forecasts from other stochastic models as well as with forecasts from a so-called conceptual hydrological model described in Section 18.3.3 (i.e., a model which attempts to mathematically simulate the physical processes involved in the hydrological cycle). It is concluded that the TFN model is the preferred model for forecasting the quarter-monthly Lac St. Jean inflow series.

18.3.2 Constructing the Time Series Models

The application involves a series of quarter-monthly natural inflows in m^3/s to the Lac St. Jean reservoir in the Province of Quebec. One of the covariate series selected for possible incorporation in a dynamic model of the inflow was rainfall. The quarter-monthly rainfall series in mm/day represents the spatial average of rainfall over the entire 57,000 km^2 watershed (Thompstone, 1983). The second covariate series was a rather novel quarter-monthly snowmelt series in mm/day, and it was calculated using logic extracted directly from the conceptual hydrologic model which is described in the next subsection. Data were available for the years 1953-82 (Thompstone et al., 1980) but only the years 1953-79 were used in fitting the models described in this section. The other three years were reserved for the split-sample forecasting experiment described in Section 18.3.4.

Following Section 13.3, the identification, estimation and diagnostic checking stages of model construction were used to build a deseasonalized ARMA model for the Lac St. Jean inflow series. Several models were examined and, based on the AIC, the standard errors (SE's) of estimation of the model parameters, and the results of diagnostic checking, the following ARMA (3,1) model was chosen:

$$(1 - 1.430B + 0.626B^2 - 0.113B^3)z_t^{(\lambda)} = (1 - 0.653B)a_t \qquad [18.3.1]$$

where $\lambda = 0.0$ indicates the given monthly series is transformed by taking natural logarithms as in [3.4.30], the z_t inflow series is deseasonalized by subtracting seasonal means and dividing by seasonal standard deviations, and a_t is the approximately normally distributed white noise innovation having a mean of zero and a variance of $\hat{\sigma}_a = 0.685$. All AR and MA parameters were more than two SE's from zero, and thus are statistically significant. Diagnostic checking of the residuals confirmed them to be uncorrelated, homoscedastic and approximately normally distributed (see Chapter 7). The AIC of the model was found to have a value of 13,771.24.

Both the empirical approach and the Box and Jenkins procedure were used to identify TFN models (see Section 17.3.1) for forecasting Lac St. Jean inflows using first the rainfall series, then the snowmelt series, and then both series together. The rainfall series was deseasonalized by subtracting the seasonal mean from each observation, and then dividing this by the seasonal standard deviation. The sample autocorrelation function (ACF) calculated using [2.5.9] showed the resulting series to be white noise. The sample CCF (cross correlation function) between the deseasonalized rainfall and the deseasonalized, logarithmic inflow series is shown in Figure 18.3.1. The 95% confidence limits in this figure are calculated under the assumption that the sample CCF values are NID$(0,n^{-1})$ where n is the length of the series (see Section 16.2.2). Because riverflows are caused by rainfall, the values of the sample CCF are significantly large for zero and negative values in Figure 18.3.1. As a result of the extra large value at lag -1, the order of the operator in the numerator of the transfer function in [18.2.2] is $m = 1$. The dying out effect for negative lags suggests that $r = 1$ for the operator in the denominator of the TFN in [18.2.2]. This form of model was fit to the data, and the resulting noise was identified as being ARMA(2,1). Consequently, the TFN model which gives the relationship between deseasonalized rainfall, x_{t1}, and deseasonalized logarithmic inflow, y_t, was selected to be:

$$y_t = \frac{(\omega_0 - \omega_1 B)}{(1 - \delta_1 B)} x_{t1} + N_t \qquad [18.3.2]$$

where

$$N_t = \frac{(1 - \theta_1 B)}{(1 - \phi_1 B - \phi_2 B^2)}$$

Table 18.3.1 provides the MLE's (maximum likelihood estimates) of the parameters and their corresponding SE's. Diagnostic checking showed the residuals to be uncorrelated and approximately normally distributed. The AIC and residual standard deviation for the rainfall and inflow TFN model were found to be 13,159.76 and 0.583, respectively. These values compare with 13,771.24 and 0.685, respectively, for the deseasonalized inflow ARMA(3,1) model. Thus, the inclusion of the rainfall series into the modelling has improved the accuracy of the model for the inflow series.

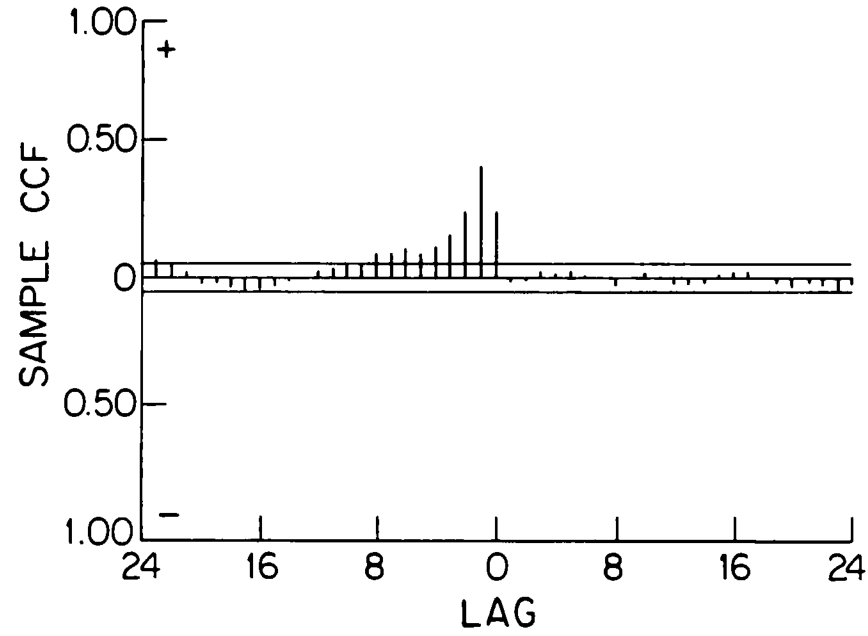


Figure 18.3.1. Sample CCF between deseasonalized, logarithmic inflow and deseasonalized rainfall series along with the 95% confidence limits.

Table 18.3.1. Parameter estimates and SE's for rainfall to inflow TFN model.

Parameters	MLE's	SE's
δ_1	0.608	0.034
ω_0	0.257	0.016
ω_1	-0.277	0.018
ϕ_1	1.410	0.173
ϕ_2	-0.472	0.127
θ_1	0.762	0.163

For the case of the snowmelt series, the selected deseasonalization involved only the subtraction of the seasonal mean from each observation. The following AR(2) model was identified and fitted to the deseasonalized series:

$$(1-\phi_{1,2}B-\phi_{2,2}B^2)x_{t2}=a_{t2} \qquad [18.3.3]$$

where the estimates of the parameters and their corresponding of SE's are given in Table 18.3.2. Diagnostic checking showed that the residuals were uncorrelated and approximately normally distributed.

Table 18.3.2. Parameter estimates and SE's for the AR(2) model of snowmelt.

Parameters	MLE's	SE's
$\phi_{1,2}$	0.267	0.027
$\phi_{2,2}$	-0.156	0.027

In accordance with the Box and Jenkins identification procedure of Section 17.3.1, the y_t output series was filtered, using [17.3.5] to obtain the estimated AR operator in [18.3.3], to produce the filtered output

$$\hat{\beta}_t = (1 - 0.267B + 0.156B^2)y_t \qquad [18.3.4]$$

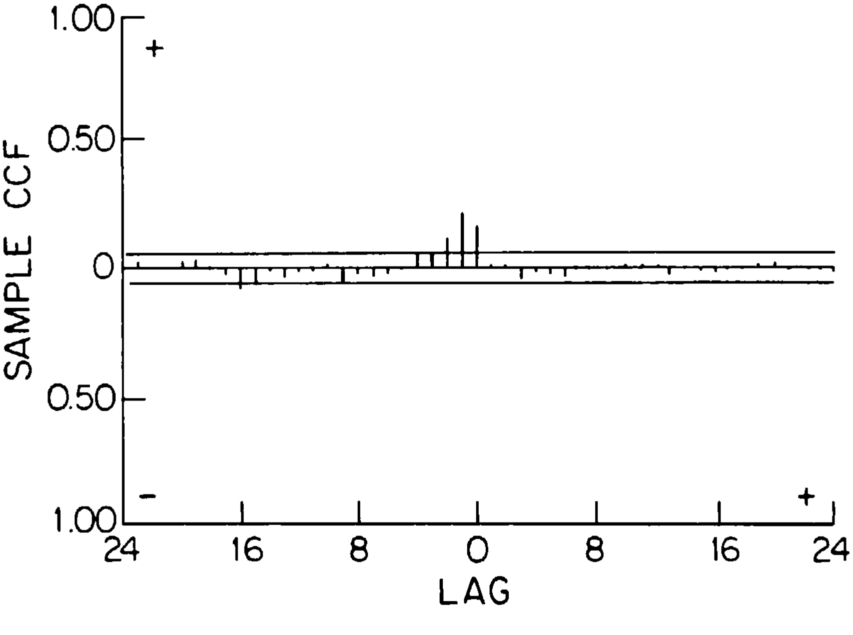


Figure 18.3.2. Sample CCF between the filtered inflow and prewhitened deseasonalized snowmelt.

The sample CCF between the prewhitened deseasonalized snowmelt series and the transformed output series is shown in Figure 18.3.2. This CCF suggested that the form of the transfer function be $r = 1$ and $m = 1$ in [18.2.2]. Such a model was fitted to the data, and the remaining noise was identified as AR(1). The TFN model chosen to relate deseasonalized snowmelt, x_{t2}, and deseasonalized logarithmic inflow, y_t, was therefore:

$$y_t = \frac{(\omega_0 - \omega_1 B)}{(1 - \delta_1 B)} x_{t2} + N_t \qquad [18.3.5]$$

where

$$N_t = \frac{a_t}{(1 - \phi_1 B)}$$

and the estimates of the parameters and their SE's are as given in Table 18.3.3. The residuals were shown to be independent and approximately normally distributed. The AIC and residual standard deviation for this snowmelt to inflow TFN model were found to be 13,495.36 and 0.664, respectively. These results suggest that a model of inflows including the relationship with snowmelt is better than a model without snowmelt, but that the rainfall series is of more use than the snowmelt series in explaining inflow.

Table 18.3.3. Parameter estimates and SE's for snowmelt to the inflow TFN model.

Parameters	MLE's	SE's
δ_1	0.541	0.098
ω_0	0.113	0.015
ω_1	-0.083	0.018
ϕ_1	0.717	0.019

In order to further improve the modelling of the Lac St. Jean inflows, a TFN model including both the rainfall and snowmelt covariate series was constructed. The form of the transfer functions in [18.3.2] and [18.3.5] was conserved (i.e., $r = m = 1$), a model was estimated, and an ARMA(2,1) model was identified for the resulting noise series. The final model for explaining the deseasonalized logarithmic inflow series, y_t, as a function of the deseasonalized rainfall, x_{t1}), and snowmelt, x_{t2}, series was thus:

$$y_t = \frac{(\omega_{0,1} - \omega_{1,1} B)}{(1 - \delta_{1,1} B)} x_{t1} + \frac{(\omega_{0,2} - \omega_{1,2} B)}{(1 - \delta_{1,2} B)} x_{t2} + N_t \qquad [18.3.6]$$

where

$$N_t = \frac{(1 - \theta_1 B)}{(1 - \phi_1 B - \phi_2 B^2)} a_t$$

The estimates of the parameters and their SE's are given in Table 18.3.4 for the transfer functions, and in Table 18.3.5 for the noise term.

Table 18.3.4. Parameter estimates and SE's (in brackets) of estimates for the transfer functions in [18.3.6].

Series	j	$\delta_{1,j}$	$\omega_{0,j}$	$\omega_{1,j}$
Deseasonalized Rainfall	1	0.625 (0.033)	0.233 (0.016)	-0.269 (0.018)
Deseasonalized Snowmelt	2	0.579 (0.090)	0.102 (0.013)	-0.046 (0.017)

Note: Parenthetical figure is SE of estimation.

Diagnostic checking of the residuals from the fitted model in [18.3.6] suggested they were normally distributed. Figure 18.3.3 shows a plot of the values of the residual autocorrelation function (RACF) and their 95% confidence intervals, defined in Section 7.3.2. Because all of the values of RACF except one fall within the 95% confidence limits, the residuals are white. The large value at lag 26 is probably due to chance and not the lack of a suitable model. Further diagnostic checking involved cross correlation functions. Figure 18.3.4 shows the cross correlations between the deseasonalized rainfall series and the residuals for the TFN model in [18.3.6], while Figure 18.3.5 shows the values of the CCF between residuals of the AR(2) deseasonalized snowmelt series in [18.3.3] and the residuals of [18.3.6]. Because the values of the CCF in Figures 18.3.4 and 18.3.5 fall within the 95% confidence interval, the noise term in the TFN model is not correlated with the prewhitened input series.

The AIC for the TFN model in [18.3.6] is 13,074.37, and the residual standard deviation is 0.562. These two measures confirm that the use of both the rainfall and snowmelt covariate series better explains the inflow series than the employment of either of the series individually. Table 18.3.6 provides a summary comparison of the AIC values and the residual standard deviations of the four models of the Lac St. Jean uncontrolled inflows developed in this section. Note that it can be shown theoretically that the MMSE forecasts from the TFN model of [18.3.6] are more accurate than those from the deseasonalized ARMA model. This fact is confirmed by the forecasting experiment described in Section 18.3.4.

Finally, in Section 14.6 a PAR model was fitted to the Lac St. Jean quarter-monthly inflow series. The AIC of this model was calculated as 13,681.61, and this suggested it was preferable to the deseasonalized ARMA model, but not as good as any of the TFN models. Nevertheless it was retained for use in the forecasting experiment described in Section 18.3.4.

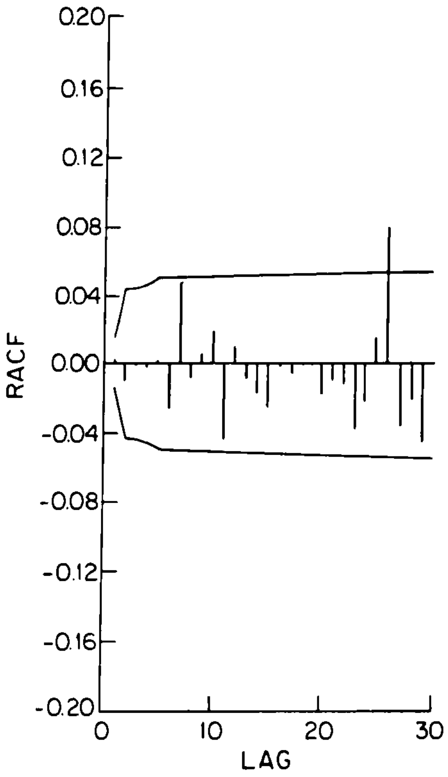


Figure 18.3.3. RACF and 95% confidence interval for the TFN model in [18.3.6].

Table 18.3.5. Parameter estimates and SE's for the noise model in [18.3.6].

Parameters	MLE's	SE's
ϕ_1	1.311	0.123
ϕ_2	-0.382	0.092
θ_1	0.712	0.113

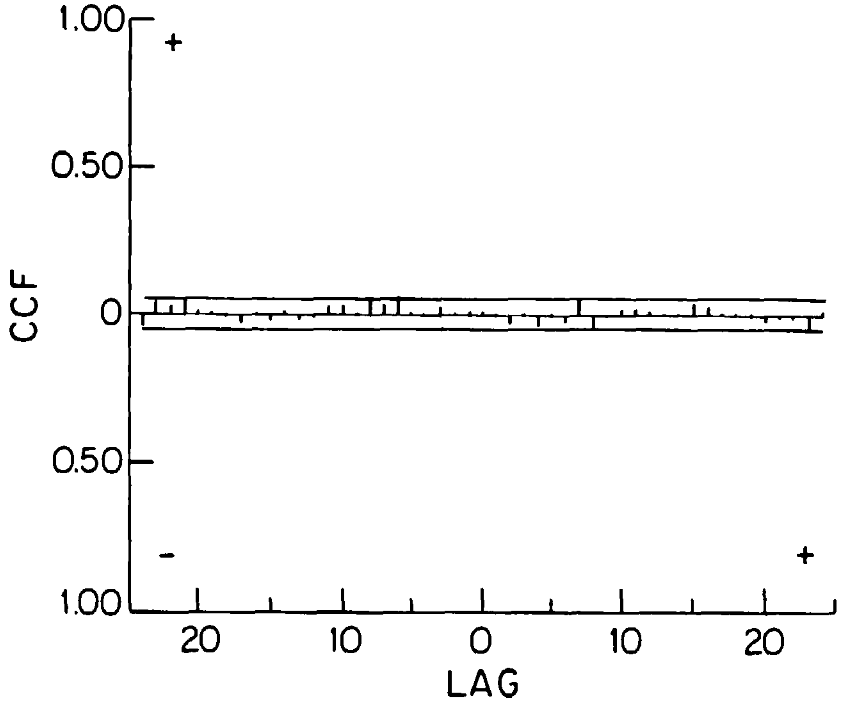


Figure 18.3.4. CCF between the deseasonalized rainfall and residuals of the TFN model in [18.3.6] along with the 95% confidence interval.

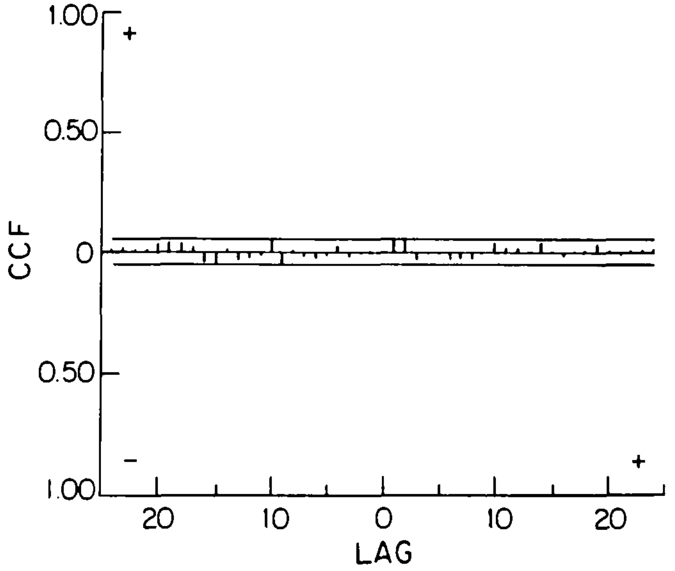


Figure 18.3.5. CCF between residuals of AR(2) deseasonalized snowmelt series and residuals of the TFN model in [18.3.6] along with the 95% confidence interval.

Table 18.3.6. Comparisons of AIC and $\hat{\sigma}_a$ values for the deseasonalized ARMA and TFN models.

Input Series	m	s	Noise	AIC	$\hat{\sigma}_a$
-	-	-	(3,1)	13,771.24	0.685
Deseasonalized Rainfall	1	1	(2,1)	13,159.76	0.583
Deseasonalized Snowmelt	1	1	(1,0)	13,495.36	0.664
Deseasonalized Rainfall Deseasonalized Snowmelt	1 1	1 1	(2,1) ·	13,074.37 ·	0.562 ·

18.3.3 Conceptual Hydrological Model

A realtime daily hydrological forecasting system has been developed for use in the operational management of the hydroelectric system operated by Alcan Smelters and Chemicals Ltd., in the Saguenay-Lac St. Jean region of Quebec. The forecasting system (Thompstone et al., 1981) uses a lumped parametric conceptual hydrological model to simulate the relationship between daily meteorological conditions and natural inflows to various reservoirs. When the forecasting system is executed, recent meteorological conditions are represented using recent measurements at meteorological stations, and future meteorological conditions are represented using meteorological forecasts provided by the Atmospheric Environment Service of Environment Canada. The basic strategy in the selection of a conceptual hydrological model was to choose a simple and flexible model in preference to more elaborate models, provided no significant improvement in the accuracy of the forecasts could be obtained by the more complex models.

There exists a multitude of conceptual models which have been used in operational hydrological forecasting, each model having its particular strengths and weaknesses (World Meteorological Organization, 1975). The conceptual model chosen for the Alcan forecasting system was originally developed by S.I. Solomon and Associates (1974), and subsequently modified by Kite (1978), the modified model being called the Water Resources Branch model. It has undergone further modifications since inclusion in the Alcan system. A detailed description of the model and the reasons it was chosen are contained in Thompstone (1983) and references therein.

The realtime daily hydrological forecasting system which uses the conceptual hydrological model provides hydrological forecasts based on meteorological forecasts and long term daily meteorological statistics (Thompstone et al., 1981). This system, referred to as PREVIS, has been operational since March, 1979, and it can be executed on a daily basis to provide hydrological forecasts for seven days into the future. The meteorological forecasts have been obtained, interpreted and entered into the forecasting system only on weekdays. Consequently, meteorological forecasts were not available for use in the proposed forecasting study.

In order to provide a basis for comparison of forecasts from the conceptual hydrological model, it was decided that observed meteorological conditions would be used in place of the meteorological forecasts and long term statistics. In other words, the conceptual hydrological model was used in the simulation mode rather than the forecasting mode. Thus, results of the forecasting study are biased in favour of the forecasts from the PREVIS system.

In using the PREVIS system, it has been recognized that the model generally follows the trends of inflows, but during certain periods is consistently higher or lower than the observed inflows. Consequently, an ad hoc smoothing of the raw hydrological forecasts was introduced into the system. The inflow forecast for the next seven days is adjusted by adding the average error of simulated versus observed inflows for the previous seven days. During the spring period, since inflows vary relatively rapidly, the smoothing period is reduced to the previous three days. In order to approximate this crude smoothing, a second set of so called forecasts from the PREVIS system was developed by adjusting the inflow forecast for the next quarter-month period by the error for the previous quarter-month period. These forecasts are labelled herein as PREVIS/S.

Note that in order to compare forecasts from the PREVIS and PREVIS/S models in the same domain as forecasts from the other models, these former forecasts are transformed using natural logarithms. This is necessary since the Pitman (1939) correlation test (see Section 8.3.2) used to compare mean squared errors of forecasts is based on the forecast errors being approximately normally distributed.

18.3.4 Forecasting Experiments

In order to compare the forecasting abilities of the deseasonalized ARMA models, TFN model, PAR model and conceptual hydrological model, a split-sample approach was adopted whereby one-step-ahead quarter-monthly forecasts were generated for three years of data, from the beginning of 1980 to the end of 1982. Data from these years were not used in either building the time series models or in calibrating the conceptual hydrological model.

Until recently, a great deal of effort had been devoted to the advancement of forecasting procedures while relatively little research had been devoted to developing methods for evaluating the relative accuracy of the forecasts produced by the different procedures (Thompstone et al., 1985; Noakes et al., 1985, 1988). Granger and Newbold (1973, 1977) have provided useful comments concerning the evaluation of forecasts and the costs of errors. The mean square error (MSE) is a cost function which is intuitively simple to understand and has been widely used in previous forecasting studies. It is the MSE and its square root, the standard error of forecast, which are used herein to compare the competing forecasting models. Various forecasting tests are discussed in detail in Section 8.3.2 and utilized in forecasting experiments carried out in Chapters 8, 15 and 18.

Noakes et al. (1985, 1988) have underlined the importance of not simply ranking models according to the MSE's of competing procedures. In their study, they used the test of Pitman (1939) and a likelihood ratio test as well as a nonparametric test to compare the one-step-ahead forecasts from different models (see also Sections 8.3.4, 15.3 and 15.4). Since the tests led to essentially the same conclusions, and the Pitman test is computationally less demanding, it has been adopted for the current research.

In order to describe the Pitman (1939) test, which is also presented in Section 8.3.2, let $e_{1,t}$ and $e_{2,t}$ $(t = 1,2, \ldots ,L)$ denote the one-step-ahead forecast errors for models 1 and 2 respectively. Then, the null hypothesis from [8.3.2] is

$$H_o: MSE(e_{1,t}) = MSE(e_{2,t}) \qquad [18.3.7]$$

where $MSE(e) = <e^2>$, and $<.>$ denotes expectation. The alternative hypothesis, H_1, is the negation of H_0.

As explained in Section 8.3.2 just after [8.3.2], for Pitman's test, let $S_t = e_{1,t} + e_{2,t}$ and $D_t = e_{1,t} - e_{2,t}$. Pitman's test is equivalent to testing if the correlation, r, between S_t and D_t is significantly different from zero. Therefore, provided $L > 25$, H_0 is significant at the 5% level if $|r| > 1.96/\sqrt{L}$.

The results of the forecasting study are summarized in Tables 18.3.7 and 18.3.8. Table 18.3.8 shows the root mean squared errors (RMSE's) of the forecasts of the logarithmic series for the five different models. The model with the smallest RMSE is the TFN model, while the second best model is the deseasonalized ARMA model. The worst forecasts are provided by the PREVIS model, while the PREVIS/S model is second worst.

Table 18.3.7. RMSE's of forecasts for the logarithmic quarter-monthly Lac St. Jean uncontrolled inflows from 1980 to 1982.

Models	RMSE's
ARMA/DES	0.298
PAR	0.301
PREVIS	0.389
PREVIS/S	0.354
TFN	0.278

Table 18.3.8. Correlation test statistics and forecast errors for the forecasts for the quarter-monthly logged Lac St. Jean uncontrolled inflows from 1980 to 1982.

Models	ARMA/DES	PAR	PREVIS	PREVIS/S	TFN
ARMA/DES*		0.0296 (=)	0.2675 (+)	0.1814 (+)	0.0902 (=)
PAR*	0.0296 (=)		0.2561 (+)	0.1704 (+)	0.1000 (=)
PREVIS	0.2675 (-)	0.2561 (-)		0.0995 (=)	0.3225 (-)
PREVIS/S	0.1814 (-)	0.1704 (-)	0.995 (=)		0.2421 (-)
TFN	0.902 (=)	0.1000 (+)	0.2421 (+)	0.2421(+)	

(1) Table shows $|r|$.

(2) Difference in MSE's of forecasts significant at 5% level if $|r| > 0.163$.

(3) A parenthetical = indicates the difference is not significant, a + indicates the row model is "better" than the column model (significant difference and smaller MSE), and a – indicates the row model is "worse" than the column model.

(4) * indicates the model is better or equal to all other models.

Table 18.3.8 examines the statistical significance of differences in the mean squared errors of forecasts from the various models. Using a 5% significance level, it is concluded that each of the time series models is better than or equal to the PREVIS and PREVIS/S models. There is no significant difference in forecasts from the ARMA/DES, PAR, and TFN models. However, since the TFN model has the smallest RMSE of forecasts and is favoured with respect to the AIC and residual variance, it is recommended that it be adopted for forecasting the Lac St. Jean inflow on a quarter-monthly basis. The physical relationship known to exist between rainfall, snowmelt and inflow reinforces this recommendation. Furthermore, a comparison of the RMSE's of forecasts in the inflow domain for which the flows are not logarithmic confirms that forecasts from the TFN model are preferable to forecasts from the conceptual model (RMSE of 512.30 as opposed to 625.85).

18.3.5 Conclusions

The TFN model described in [18.3.6] provides an effective means of forecasting quarter-monthly inflows to the Lac St. Jean reservoir based on rainfall and snowmelt. The most recent statistical techniques and an understanding of the physical processes involved are used to identify, estimate and verify a reasonable model. Both the empirical approach and the Box and Jenkins approach are useful in model identification (see Section 17.3.1). The MAICE procedure (Section 6.3) indicates the TFN model with both covariate series is better than a deseasonalized ARMA model, PAR model or TFN model with only one or another of the covariate series. The split-sample forecasting experiments of Section 18.3.4 demonstrate that the full TFN model provides better forecasts than a particular conceptual hydrological model. Consequently, the TFN model is the preferred model for forecasting the quarter-monthly Lac St. Jean inflow series. It is interesting to note that Chow et al. (1983) also found flood forecasts from a TFN model to be as reliable as forecasts generated from a complex conceptual model. Hence, they concluded that TFN models provide an attractive alternative to conceptual models for use in realtime flood forecasting.

18.4 COMBINING HYDROLOGICAL FORECASTS

18.4.1 Overview

Often a variety of models can be fitted to a given data set. For example, in Section 18.3, time series models consisting of TFN, PAR (Chapter 14) and deseasonalized ARMA (Chapter 13) models, plus two related conceptual models, are fitted to a hydrological time series. Each of these calibrated models can then be employed for generating forecasts for the series. Although one model may produce more accurate forecasts than others in the long run, it may not do so in every instance. Consequently, one may wish to improve the forecasts by combining forecasts from two or more models in accordance to their relative performances.

The objective of this section is to show how better forecasts can be obtained when TFN forecasts are combined with other types of forecasts. In particular, a TFN, PAR and two similar conceptual models are employed to forecast quarter monthly riverflows, as is done in Section 18.3. These models all approach the modelling and forecasting problem from three different perspectives and each has its own particular strengths and weaknesses. The forecasts generated by the individual models are combined in an effort to exploit the strengths of each model. The results of this case study indicate that significantly better forecasts can be obtained when forecasts from different types of models are combined. In particular, the best forecasts are obtained when TFN and PAR forecasts are optimally combined. These forecasting experiments are also reported by McLeod et al. (1987).

Formulae for combining forecasts in an optimal manner from competing models are presented in Section 15.5.2. Additionally, forecasting experiments are presented in Section 15.5.3 for combining forecasts for monthly riverflows using SARIMA (Chapter 12) and PAR (Chapter 14) models. Because the SARIMA model is not well designed for modelling monthly riverflows for which there is stationarity within each season (see the introduction to Part VI and Section 12.1), combining forecasts from this model with the better forecasts from the PAR model does not produce improved forecasts.

18.4.2 Combination Forecasting Experiments

The data used in this section are identical to those employed in the forecasting experiments of Section 18.3. More specifically, the quarter-monthly inflows for the Lac St. Jean reservoir are utilized. Recall from Section 18.3.1, that accurate quarterly-monthly forecasts for riverflows are required so that Alcan can optimally generate hydro-electrical power for use in its aluminum smelters.

Thirty years of quarter-monthly riverflows are available from 1953 to 1982, inclusive. As is done in Sections 18.3.2 and 18.3.3, models are fitted to the first twenty-seven years of the data and then used to forecast the one-step-ahead forecasts for the last three years. Prior to fitting models to the riverflows, the data are first transformed using natural logarithms.

The calibrated models used in the study are already described in Sections 18.3.2 and 18.3.3. In particular, the finite difference equation for the best TFN model is given in [18.3.6] while its parameter estimates are listed in Tables 18.3.4 and 18.3.5. The most appropriate PAR model is identified using graphs of the sample periodic ACF and PACF (defined in Section 14.3.2). The two versions of the conceptual model used in the combination forecasting study are the PREVIS and PREVIS/S conceptual models described in Section 18.3.3.

The RMSE's of the logarithmic forecast errors are presented in Table 18.3.7. As can be seen, the TFN model has the smallest RMSE of all the models considered. As such, this value will be used as a basis for comparison of the various techniques employed to combine the individual forecasts.

Notice that the deseasonalized ARMA and PAR models have almost the same RMSE's in Table 18.3.7. Because the PAR model is generally better to use than the deseasonalized model for modelling seasonal riverflows for which there are sufficient data (see discussion in Part VI), the deseasonalized ARMA model is not employed in the combination experiments of this section.

The equations for combining forecasts are given in Section 15.5.2. In this study, the weights for combining the individual forecasts were calculated using both [15.5.2] and [15.5.4] with $\upsilon = 4$, 8 and 12. Since the model residuals were not employed, the first υ forecasts were combined using equal weights. The weights were then recalculated for each subsequent forecast using the previous υ forecast errors.

The forecasts from the four models were combined in a pairwise fashion with the exception of the two conceptual models (PREVIS and PREVIS/S). The resulting RMSE's of the combined forecasts using [15.5.2] to calculate the combining weights are given in Table 18.4.1. The subscripts associated with the RMSE's indicate the number of previous forecast errors that were employed to calculate the weights. For example, when the previous four forecast errors were used to combined the TFN and PAR forecasts, the resulting RMSE was 0.142. In most cases, the greater the number of previous forecast errors employed to calculate the weights, the smaller the resulting combined RMSE.

Table 18.4.1. RMSE's of the combined quarter monthly forecasts with combining weights calculated using [15.6.2].

Model Combinations	$RMSE_4$	$RMSE_8$	$RMSE_{12}$
TFN - PAR	0.142	0.120	0.119
TFN - PREVIS	0.787	0.524	0.418
TFN - PREVIS/S	0.994	0.318	0.271
PAR - PREVIS	0.243	0.229	0.222
PAR - PREVIS/S	0.217	0.186	0.187

The smallest RMSE was obtained when the TFN and PAR forecasts were combined using the previous 12 forecast errors to calculate the weights. The resulting RMSE was less than half the value of the smallest RMSE for the individual models suggesting that significant benefits can be obtained by combining the forecasts from these two models. Conversely, the largest RMSE's were found when the TFN forecasts were combined with the PREVIS or PREVIS/S forecasts. Only when the previous 12 forecast errors were employed to calculate the weights did the combined TFN and PREVIS/S forecasts yield a smaller RMSE than the best individual model. Even then, the difference was only in the third decimal place.

The resulting RMSE's of the combined forecasts when [15.5.4] was employed to calculate the combining weights are given in Table 18.4.2. In this case, only one combination had a larger RMSE than the best individual model. Once again, the smallest RMSE was found when the TFN and PAR forecasts were combined using the previous 12 forecast errors to calculate the combining weights. The largest RMSE's were found when the TFN forecasts were combined with the forecasts from the two conceptual models. These RMSE's did, however, represent a significant improvement when compared to the RMSE's obtained when [15.5.2] was used to calculate the combining weights. In the previous case, poor estimates of Σ in [15.5.3] resulted in the calculation of one negative weight and one weight greater than one. As a result, the corresponding RMSE's were more than three times as large as the RMSE of the best individual model. It is therefore recommended that, unless reasonably good estimates of Σ can be obtained, the suboptimal estimates of the combining weights calculated using [15.5.2] be employed.

As a test of combining forecasts from more than two models, the forecasts produced by the TFN, PAR and PREVIS/S models were combined using equal weights. The resulting RMSE was 0.136. Although this does not represent the lowest RMSE, even this naive combination of forecasts produced a RMSE which was less than half the RMSE of the best individual model.

Table 18.4.2. RMSE's of the combined quarter-monthly forecasts with combining weights calculated using [15.5.4].

Model Combinations		$RMSE_4$	$RMSE_8$	$RMSE_{12}$
TFN	- PAR/PACF	0.146	0.124	0.122
TFN	- PREVIS	0.275	0.251	0.247
TFN	- PREVIS/S	0.283*	0.252	0.250
PAR/PACF	- PREVIS	0.244	0.230	0.222
PAR/PACF	- PREVIS/S	0.214	0.187	0.188

*Larger RMSE than TFN forecast error in Table 18.3.7.

18.4.3 Conclusions

Combining economic forecasts from various models has become fairly common practice. However, the case studies presented in Sections 18.4.2 and 15.5.2 as well as by McLeod et al. (1987) represent the first reported experiments dealing with the combination of riverflow forecasts. Combining forecasts from conceptual models, a TFN model and a PAR model resulted in a significant reduction in the RMSE's of the forecasts. These three models approach the modelling problem from three distinctly different perspectives. The relative strengths of each model were enhanced by combining the individual forecasts. Thus, based upon the results of this case study, it would appear that significant improvements in forecasting performance can be obtained when the forecasts from different types of models are combined.

18.5 RECORD EXTENSIONS, CONTROL AND SIMULATION

18.5.1 Overview

The main objectives of this chapter are to explain how reliable forecasts can be calculated using TFN models and to demonstrate how forecasting can be conveniently carried out in practice using the hydrological forecasting experiments of Sections 18.2.3, 18.3 and 18.4. The purpose of this section is to outline how TFN models can be employed for three other kinds of applications: extensions of hydrologic records, control and simulation.

18.5.2 Record Extensions

Using natural time series records from the Arctic, Baracos et al. (1981) explain how hydrometric records can be extended using TFN models. In particular, weather records have been kept in the Arctic for a much longer period of time than have hydrometric or riverflow measurements. Based on a knowledge of the dynamic relationship between riverflow series and meteorologic series, it is possible to give an estimate of the values the hydrometric series is likely to have taken during the period when weather data are available, but before flow records were kept. This may be thought of as an artificial extension of the hydrometric record and can be considered to be a type of *back forecasting*. The true values of the unmeasured flows can of course never be obtained by this method, but likely values, given the covariate meteorologic input series, can be calculated. These estimates are simply the output of the TFN model with the noise term set to its conditional expectation of zero.

Baracos et al. (1981) develop ARMA, TFN and intervention models (see Chapter 19) for modelling 16 average monthly riverflow series as well as precipitation and temperature series from the Northwest Territories in the Canadian Arctic. The data sets are available from the Water Survey of Canada which is part of Environment Canada in Ottawa. To explain how riverflow records can be extended using meteorological inputs, consider the TFN model developed for the flows of the Tree River. Average monthly flows for the Tree River are available for 8 years from the start of 1969 to the end of 1976. However, the two meteorologic input series consisting of precipitation and temperatures from the Coppermine weather station are 44 years in length and span the years from the start of 1933 to the end of 1976. For the years in which the riverflows overlap with the meteorologic data, a TFN model can be developed to model how the meteorologic inputs dynamically affect the riverflow output series. The TFN model can then be employed for extending or back forecasting the riverflow series for the years during which there are only meteorological records.

The calibrated TFN model for the Tree River is written as

$$y_t = 0.0012x_{t1} + 0.04x_{t2} - 0.031Bx_{t3} + \frac{1 - 0.32B + 0.25B^8}{1 + 0.57B}a_t \qquad [18.5.1]$$

where

y_t is the Tree River series which is first transformed by taking natural logarithms and then deseasonalized by removing the monthly means for the logarithmic series using [13.2.2].

x_{t1} is the Coppermine rainfall series which is deseasonalized by subtracting the appropriate monthly mean from each observation. Snowmelt is included as part of the rainfall series. In order to produce a plausible representation of snowmelt input to a riverflow series, the monthly snowfalls are summed over each winter, and then the total snowfall for the winter is introduced as a pulse input to the rainfall series during the first month that the mean temperature rises above zero Celsius for each year. Snowfalls that occur during months when the mean temperature is above zero Celsius are assumed to have melted immediately, and are added to the rainfall series rather than to the winter's snow accumulation.

x_{t2} is the Coppermine temperature series which is deseasonalized by removing monthly means. Because the temperature is below zero in the winter and hence does not melt the snow, the values from January, February, March, November and December are set to zero.

x_{t3} is input series containing the deseasonalized temperature only for the month of April. All other months are set equal to zero. The reason for including the x_{t3} series in the third term on the right hand side of [18.5.1] is because for the month of April there is a large negative cross correlation at lag one between the prewhitened Tree riverflows and the Coppermine temperature series.

a_t is the noise term for the TFN model which is NID$(0,\sigma_a^2)$.

To employ the calibrated TFN model for extending the riverflows, the conditional expectation of the noise is assumed to be zero and hence one uses only the dynamic component on the right hand side of [18.5.1] to calculate y_t as

$$y_t = 0.0012x_{t1} + 0.04x_{t2} - 0.031Bx_{t3} \qquad [18.5.2]$$

By substituting in known values of x_{t1}, x_{t2} and x_{t3} in [18.5.2], one can determine y_t's for any desired values of t. Subsequently, to find the values of the flows in the original untransformed domain one simply takes the inverse deseasonalization and logarithmic transformation of the generated y_t's from [18.5.2].

Using the above procedure, the average monthly flows for the Tree River can be predicted for any period during the years for which meteorological records exist from 1933 to 1976. By utilizing graphical and numerical results, Baracos et al. (1981) demonstrate that the predicted flows using [18.5.2] produce reasonable results. In particular, during the time period for which the flows are known, from 1969 to 1976, the predicted flows are close to the known historical flows.

Following a similar procedure to the one described in this section, Beauchamp et al. (1989) extend daily riverflow records of a downstream station based upon a TFN noise model that connects the downstream flows to a longer upstream time series of daily riverflows. They also employ regression analysis for extending the same riverflows. However, they point out that the regression model was found to have a significant amount of correlation in the residuals which the TFN could eliminate, since the noise in a TFN model can be modelled as an ARIMA model.

Snorrason (1986) employs a TFN model to extend seasonal riverflow records for a river in Iceland. A longer temperature series constitutes the input to the TFN model which has the riverflows as the output.

18.5.3 Control

This chapter deals mainly with employing TFN models for forecasting or predicting the future values of the response variable. As pointed out by Young (1984, p. 104), another important application area of TFN models is designing control and management schemes for the system that is currently being studied. In the chemical industry, for example, TFN models are employed extensively for scientifically controlling processes for optimally producing a wide range of chemical products. The key reason why TFN models are ideally suited for control purposes is that they mathematically describe how the inputs dynamically affect the response in the presence of correlated noise.

In a control problem, one often wishes to keep a response variable as close as possible to a target value in a system subject to the inputs and noise. One could attempt to design control schemes which minimize an overall measure of error at the output such as the mean square error. As explained by Box and Jenkins (1976, Chapter 12), one can categorize control procedures into three main domains - feedforward control, feedback control and a mixture of these two. In *feedforward control,* one or more sources of disturbances (inputs) are measured and these observations can be employed for compensating for potential deviations in the output. Because input into the system is used to control the output of the system, this is referred to as feedforward control. On the other hand, in some applications the only information available about the existence of the input disturbances is the deviation from the target which they cause in the response. If only this deviation is utilized for deciding upon how to adjust the system, the action is called *feedback control.* A combination of the aforementioned two methods of control is referred to as *feedforward-feedback control.*

For a detailed discussion of discrete control schemes, the reader may wish to refer to Part IV of Box and Jenkins (1976). Certainly, the design of control schemes has many potential applications in water resources and environmental engineering. For example, to maximize the hydroelectrical output of a system of reservoirs, good control and management plans are required. The efficient operation of a sewage treatment facility that handles both industrial and residential liquid wastes poses many interesting control problems.

18.5.4 Simulation

Besides forecasting, a TFN model can, of course, also be employed for *simulation* purposes. To simulate with a TFN model, it is most convenient to use the model as given in [18.2.13] or [18.2.20] where appropriate multiplications have been made so that no operators appear in the denominator in any term on both sides of the equation. For explanation purposes, consider the TFN model having one input series and ARMA noise which is written in [18.2.13] for time t as

$$\phi(B)\delta(B)(y_t-\mu_y)=\phi(B)\omega(B)(x_t-\mu_x)+\delta(B)\theta(B)a_t \qquad [18.5.3]$$

The main steps to follow in simulating with a TFN model are:

1. By employing the ARMA model that is separately fitted to the x_t series in [18.2.4], use the simulation techniques of Section 9.3 or 9.4 to simulate the x_t's.
2. To simulate the a_t's needed in the second term on the right hand side of [18.5.3], employ an appropriate method from Section 9.2.3 to simulate the a_t's which are NID$(0,\sigma_a^2)$.
3. If starting values are needed for the y_t's in [18.5.3], these can be generated using a separate ARMA model fitted to the y_t series in conjunction with a simulation technique from Section 9.3 or 9.4.
4. Use the simulated x_t and a_t series from steps 1 and 2, respectively, as well as the starting values for y_t from step 3, in the TFN model in [18.5.3] to simulate the y_t series.

18.6 CONCLUSIONS

The TFN model of Chapter 17 is particularly well designed for use in the natural sciences such as hydrology and water quality modelling. This is because the TFN model in [18.2.18] and [17.5.3] can formally describe, using a finite difference equation, the dynamic relationships existing between a single output series and one or more input series. For instance, the TFN model in [18.3.6] describes how the input or covariate series consisting of rainfall and snowmelt cause riverflows. Furthermore, the correlated noise in the model can be modelled using an ARMA(2,1) model.

Because the structure of the TFN model in an equation such as [18.3.6] realistically reflects the physical relationships among the variables, one would expect the model to provide good forecasts. In addition, since the TFN model incorporates more information into its structure by means of the input series, one would think that better forecasts should be obtained using this model. Indeed this is exactly what happens. The forecasting experiments of Section 18.3 demonstrate that the TFN noise model forecasts seasonal riverflows better than its competitors. In particular, the TFN model of [18.3.6] provides more accurate forecasts of the quarter-monthly riverflows into Lac St. Jean than the deseasonalized ARMA, PAR or either of the two conceptual models. Moreover, as shown by the forecasting results in Section 18.4, even better forecasts can be obtained when the TFN forecasts are optimally combined with those provided by the PAR model.

Forecasting experiments with a range of nonseasonal models are furnished in Section 8.3. For a description of forecasting experiments with the seasonal models of Part VII, the reader can turn to Sections 15.3 and 15.4. Experiments with combinations of forecasts from seasonal models are also given in Section 15.5.3.

In addition to handling multiple input series, the TFN model of Part VII can be expanded to take care of other situations that arise in practice. More specifically, the intervention model of Part VIII constitutes a general type of TFN model that can be used to model the effects of external interventions upon the mean level of a series, estimate missing observations and also to describe the dynamic relationships between multiple input series and a single output. Besides Part VIII, further interesting applications of intervention and TFN modelling are presented in Chapter 22.

PROBLEMS

18.1 Suppose that a TFN model is written as

$$y_t - \mu_y = \frac{\omega_0 B}{(1-\delta_1 B)}(x_t - \mu_x) + \frac{(1-\theta_1 B)}{(1-\phi_1 B)}a_t$$

For this model, carry out the following tasks:

(a) Using formulae, clearly explain how to iteratively calculate MMSE forecasts for lead times $l = 1,2, \ldots,$.

(b) Derive the formula for determining the variance of the forecast error for $\hat{y}_t(l)$.

18.2 Carry out the instructions of Problem 18.1 for the following TFN model having three input series.

$$(y_t - \mu_y) = \frac{\omega_1(B)}{\delta_1(B)}(x_{t1} - \mu_{x1}) + \frac{\omega_2(B)}{\delta_2(B)}(x_{t2} - \mu_{x2})$$

$$+ \frac{\omega_3(B)}{\delta_3(B)}(x_{t3} - \mu_{x3}) + \frac{\theta(B)}{\phi(B)}a_t$$

18.3 For the TFN model written below, execute the instructions given in Problem 18.1.

$$y_t = \frac{(\omega_0 - \omega_1 B)B^2}{(1-\delta_1 B - \delta_2 B^2)}x_t + \frac{(1-\theta_1 B)}{(1-\phi_1 B)(1-B)}a_t$$

18.4 Consider the situation where one has monthly observations for both a response variable, Y_t, and an input series, X_t. Each series is first transformed by taking natural logarithms and then deseasonalized using [13.2.3]. Next, the TFN model fitted to the resulting nonseasonal series is

$$y_t = \frac{\omega_0 - \omega_1 B}{(1 - \delta_1 B)} x_t + \frac{1}{(1 - \phi_1 B)} a_t$$

where y_t and x_t are the deseasonalized response and covariate series, respectively. By employing suitable equations, explain how to calculate MMSE forecasts for y_t and x_t as well as forecasts for the original Y_t and X_t series.

18.5 In Chapter 14, periodic models are defined for application to a single seasonal time series for which there are s seasons per year. Assuming one input series and an ARMA noise term, write down the difference equations to define a periodic TFN model. Explain the drawbacks of this type of model and how these disadvantages could be overcome.

18.6 The field of econometrics deals with the development of statistical and stochastic methods for application to economic data. Find an article in the econometrics literature where one or more leading indicators are used to forecast some aspect of the economy. Outline the procedure that is employed and explain how you think it could be improved.

18.7 Within the water resources literature, locate a paper where TFN modelling is used for forecasting. Briefly describe how the forecasting study was carried out and point out any interesting facts that you discover.

18.8 Often an overly complex model does not forecast as accurately as a much simpler time series model such as a TFN model. Explain why you think this could happen. Find a paper in a field which is of interest to you where a TFN model provides better forecasts than a more complicated model, such as a conceptual model. Describe the main findings of the paper and emphasize the most interesting results.

18.9 Fit a TFN model to a nonseasonal data set where you have a response series and one input series. Employ the calibrated model to calculate MMSE forecasts for lead times from 1 to 12. Plot the forecasts along with the 95% confidence limits.

18.10 Carry out the instructions of the previous question for two monthly time series.

18.11 Find two seasonal series designated by y_t and x_t for the response and input series, respectively. Omit the last three years of the data set and then fit SARIMA (Chapter 12), deseasonalized ARMA (Chapter 13) and PAR (Chapter 14) models to the y_t series. Also, fit a TFN model to the y_t and x_t series for which the last three years of the data are not used for calibration purposes. Following the approach of Section 18.3, determine which of the four models produces the best one step ahead MMSE forecasts of the last three years of the series.

18.12 Employ the combination methods of Section 18.4 to determine if the accuracy of the forecasts obtained using the four models in Problem 18.11 can be improved by optimally combining the forecasts. Clearly explain your findings.

18.13 Select a response series for which you have a longer record for one or more input series. Fit a TFN model to the data for the time period during which the response and input series overlap. By following the procedure of Section 18.5.2, use the TFN model to extend the response series for the time interval for which only the input data are known.

18.14 Snorrason (1986) employs a TFN model to extend seasonal riverflow data from a glaciated basin in Iceland. A longer temperature record is used as the input to the TFN model while the output is the riverflows. His record extension technique is a slightly different variation of the one presented in Section 18.5.2. Describe the data extension approach of Snorrason and compare it to the one presented in Section 18.5.2.

18.15 Using equations and diagrams, explain the feedforward, feedback and mixed control schemes put forward by Box and Jenkins (1976, Chapter 12). Describe how each of these schemes could be possibly employed for modelling a water resources or environmental system.

18.16 Fit a TFN model to a data set for which you have one input series and, of course, a single response series. Follow the procedure of Section 18.5.4 to simulate a sequence of values that has the same length as the historical series. Clearly explain all of the steps that you follow and compare a graph of the simulated y_t sequence to a plot of the historical response series.

REFERENCES

For references about causality and how to fit TFN models, the reader may wish to refer to the references in Chapters 16 and 17, respectively. Moreover, further references on forecasting are listed at the ends of Chapters 1, 8 and 15.

CONTROL

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Young, P. C. (1984). *Recursive Estimation and Time-Series Analysis: An Introduction.* Springer-Verlag, Berlin.

CONCEPTUAL MODELS

Kite, G. W. (1978). Development of a hydrologic model for a Canadian watershed. *Canadian Journal of Civil Engineering*, 5(1):126-134.

Solomon, S. I. and Associates Limited (1974). *Preliminary Analysis of Potential Remote Sensing Applications in Hydrology.* Report to Environment Canada.

Thompstone, R. M. (1983). *Topics in Hydrological Time Series Modelling*. PhD thesis, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Thompstone, R. M., Bouchard, S., Pilon, P. J. and Bergeron, R. (1981). Real-time daily hydrological forecasting for a multi-reservoir hydroelectric system. In *Proceedings of the Canadian Society for Civil Engineers Fifth National Hydrotechnical Conference*, pages 37-57, Fredericton, New Brunswick.

World Meteorological Organization (1975). Intercomparison of conceptual models used in operational hydrological forecasting. Operational Hydrology Report 7, Secretariat of the World Meteorological Organization, Geneva, Switzerland.

COMBINING FORECASTS

McLeod, A. I., Noakes, D. J., Hipel, K. W. and Thompstone, R. M. (1987). Combining hydrologic forecasts. *Journal of the Water Resources Planning and Management Division, American Society of Civil Engineers*, 113(1):29-41.

DATA SETS

Thompstone, R. M., Poire, A. and Vallee, A. (1980). A hydrometeorological information system for water resources management. *INFOR*, 18(3):258-274.

FORECASTING TESTS

Granger, C. W. J. and Newbold, P. (1973). Some comments on the evaluation of economic forecasts. *Applied Economics*, 5:35-47.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press, New York.

Noakes, D. J., Hipel, K. W., McLeod, A. I., Jimenez, J. and Yakowitz, S. (1988). Forecasting annual geophysical time series. *International Journal of Forecasting*, 4:103-115.

Noakes, D. J., McLeod, A. I. and Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting, 1:179-190.*

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31:9-12.

RECORD EXTENSION

Baracos, P. C., Hipel, K. W. and McLeod, A. I. (1981). Modelling hydrologic time series from the Arctic. *Water Resources Bulletin*, 17(3):414-422.

Beauchamp, J. J., Downing, D. J. and Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin* 25(5):961-975.

Snorrason, A. (1986). Analysis of river flow data from glaciated basin in Iceland. In: *Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, Eds. Shen, H. W., Obeysekera, J. T. B., Yevjevich, V. and DeCoursey, D. G., held at Colorado State University, Fort Collins, Colorado, July 15-17, 1985. Published by the Engineering Research Center, Colorado State University, 651-663.

TRANSFER FUNCTION-NOISE FORECASTING

Alley, W. M. (1985). Water balance models in one-month-ahead streamflow forecasting. *Water Resources Research,* 21(4):597-606.

Anselmo, V. and Ubertini, L. (1979). Transfer function-noise model applied to flow forecasting. *Hydrological Sciences Bulletin,* 24:353-359.

Chow, K. C. A., Watt, W. E. and Watts, D. G. (1983). A stochastic-dynamic model for real time flood forecasting. *Water Resources Research,* 19(3):746-752.

Fay, D. M., Watt, W. E. and Watts, D. G. (1987). A stochastic real-time spring flood forecasting system for Carman, Manitoba. *Canadian Journal of Civil Engineering,* 14(1):87-96.

Haltiner, J. P. and Salas, J. D. (1988). Short-term forecasting of snowmelt runoff using ARMAX models. *Water Resources Research* 24(5):1083-1089.

Maidment, D. R., Miaou, S. P. and Crawford, M. M. (1985). Transfer function models of daily urban water use. *Water Resources Research,* 21(4):425-432.

Noakes, D. J., Welch, D. W., Henderson, M. and Mansfield, E. (1990). A comparison of preseason forecasting methods for returns of two British Columbia sockeye salmon stocks. *North American Journal of Fisheries Management,* 10(1):46-57.

Olason, T. and Watt, W. E. (1986). Multivariate transfer function-noise model of river flow for hydropower operation. *Nordic Hydrology,* 17:185-202.

Schweigert, J. K. and Noakes, D. J. (1990). Forecasting Pacific herring (clupea harengus pallasi) recruitment from spawner abundance and environmental information. *Proceedings of the International Herring Symposium,* Anchorage Alaska, 373-387.

Snorrason, A., Newbold, P. and Maxwell, W. H. C. (1984). Multiple input transfer function-noise modeling of river flow. In Maxwell, W. H. C. and Beard, L. R., Editors, *Frontiers in Hydrology,* pages 111-126. Water Resources Publications, Littleton, Colorado.

Stocker, M. and Noakes, D. J. (1988). Evaluating forecasting procedures for predicting Pacific herring (clupea harengus pallasi) recruitment in British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences,* 45(6):928-935.

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985). Forecasting quarter-monthly riverflow. *Water Resources Bulletin,* 21(5):731-741.

PART VIII

INTERVENTION ANALYSIS

A major challenge in environmental impact assessment is to model and statistically describe the effects of both man-induced and natural **interventions** upon the mean level of a natural time series. For example, how do changes in land use such as urban growth, deforestation, reservoir construction and operation, diversion canals and other planned projects affect both water quality and riverflow patterns? In addition to altering important water quality variables like total organic carbon, phosphorous and turbidity, will specific land use changes significantly affect the stochastic characteristics of the riverflows? If a large section of a forest is destroyed by fire, will the drainage characteristics and water quality variables of the affected watersheds be significantly changed? Will pollution control programs to reduce acid rain greatly decrease the alkalinity levels in lakes and streams? To properly model, analyze and statistically describe the affects of one or more interventions on a time series, the technique of **intervention analysis** can be utilized. Indeed, as exemplified by the important applications in Chapters 19 and 22, intervention analysis constitutes one of the most flexible and comprehensive statistical tools available for use in **environmental impact assessment**.

In an intervention analysis study, an intervention model is developed for describing statistically the changes in the mean level of a time series due to either natural or man-made causes. As shown in Chapter 19, the intervention model is actually a special type of TFN (transfer function-noise) model. However, due to the great import of this model for studying pressing problems in environmental impact assessment as well as other areas, Chapters 19 and 22 of this book are devoted to describing the intervention model and using environmental applications to carefully demonstrate how it can be used in practice.

In qualitative terms, an **intervention model** can be written as

response variable = dynamic component + noise

where

dynamic component = interventions + missing data + inputs

The **response variable** consists of a single output series such as total organic carbon in a river. To model the effects of one or more interventions upon the mean level of the response variable, intervention terms can be incorporated into the **dynamic component**. An **intervention component** may be needed, for example, to ascertain how newly constructed secondary pollution control procedures at upstream sewage treatment plants affect the mean level of the total organic carbon. By designing a special kind of intervention term, the dynamic component can also be used to **estimate missing observations** in the output. The water quality series used in the applications within Chapters 19 and 22 are typical of available water quality time series where often there are missing data points. An inherent advantage of this approach to data filling is that the correlation structure of the series is automatically taken into account when the estimates for the missing data points are calculated. Finally, when there are other **input series** such as riverflows and temperature, the dynamic influence of these covariate series upon the response variable can be suitably accounted for by including suitable transfer functions in the dynamic component. As is the case for the TFN model of Part VII, the **autocorrelated noise,** which cannot be described by the dynamic component, can be adequately modelled by an appropriate ARMA or ARIMA model. Furthermore, the intervention model can be used with both seasonal and nonseasonal time series.

In Chapter 19, **the intervention model is pedagogically presented** by first describing simpler situations and then adding more complexity to the model as the chapter progresses. For example, in Section 19.2 the intervention model with only multiple interventions in the dynamic component is described whereas in Section 19.5 the complete intervention model outlined in the previous paragraph is presented. Throughout the chapter, **environmental applications** are utilized to clearly demonstrate how various kinds of intervention models can be conveniently constructed by practitioners. After detecting the presence of interventions and the times at which the interventions occur, if they are not already known, an intervention model can be built by following the usual identification, estimation and model verification stages of **model development.** To design the form of the transfer functions for the intervention terms in the dynamic component, simple **identification procedures** are introduced. In order to ascertain the parameters required in a transfer function for each input series and also the parameters needed in the noise term, techniques similar to those presented in Sections 17.3.1 and 17.5.3 can be used. Subsequent to obtaining **MLE's** (maximum likelihood estimates) for the model parameters, the adequacy of the fitted model can be verified by using suitable **diagnostic tests.** Besides using the intervention model to determine the effects of the interventions upon the mean level of the output, the intervention model can be used for other applications such as forecasting and simulation.

When dealing with environmental data, such as water quality time series, often there are many missing data points where there may be long periods of time for which no observations were taken. Additionally, there may be one or more external interventions which affect the stochastic manner in which a series behaves. In other words, environmental data are often quite "messy". The major purpose of Chapters 22 to 24 in Part X of the book is to explain clearly how intervention analysis, nonparametric tests and regression analysis, respectively, can be employed in environmental impact assessment when dealing with **messy data.** As demonstrated by water quality and quantity applications in Chapter 22, when an evenly spaced time series can be estimated efficiently from unevenly spaced observations by using an appropriate data filling technique (see Section 22.2), intervention analysis constitutes a powerful parametric procedure for rigorously modelling suspected trends.

In Part X, it is explained how the **data analysis** methodology of Tukey can be used for scientifically studying data sets by adhering to the two main steps of exploratory data analysis and confirmatory data analysis (see Chapter 22 as well as Sections 1.2.4 and 5.3.2). For discovering trends in a specified set of observations, a variety of simple, yet useful, **exploratory tools** can be utilized (see Section 22.3). To formally model trends in a series which are known in advance or else detected using exploratory data analyses, different approaches can be used at the **confirmatory data analysis** stage. In particular, the ways in which **trends** can be modelled using intervention analysis, nonparametric tests and regression analysis are described in Chapters 22 to 24, respectively.

CHAPTER 19

BUILDING INTERVENTION MODELS

19.1 INTRODUCTION

As an illustrative example of how a man-induced intervention can affect the mean level of an environmental time series, consider Figure 19.1.1 which is also displayed in Chapter 1 as Figure 1.1.1. This is a graph of 72 average monthly phosphorous data points (in milligrams per litre) from January, 1972, until December, 1977, for measurements taken by the Ontario Ministry of the Environment downstream from the Guelph sewage treatment plant located on the Speed River in the Grand River basin, Ontario, Canada. In February, 1974, a pollution abatement procedure was brought into effect by implementing conventional phosphorous treatment at the Guelph station. Notice in Figure 19.1.1 the manner in which the man-made intervention of phosphorous removal has dramatically decreased the mean level of the series after the intervention date. Furthermore, as indicated by the filled-in circles in this figure, there are missing data points both before and after the intervention date. For displaying a missing observation on the graph, the missing value is simply replaced by its monthly average across all of the months. However, estimating a missing monthly observation by a specified monthly mean may not be an accurate procedure since the autocorrelation structure inherent in the time series and the effects of the intervention are ignored. Fortunately, the technique of *intervention analysis* can be used not only to estimate the missing observations where the autocorrelation structure is automatically taken into account but also to statistically model the effects of the tertiary phosphorous treatment for reducing the mean level of the series. In Section 19.4.5, intervention analysis is employed for realistically modelling the water quality time series of Figure 19.1.1 by constructing an appropriate intervention model. The study shows that there is a 75% drop in the mean level of the series where the 95% confidence interval is from 71% to 78%. Rigorous statistical statements like this can be readily obtained by using the general and flexible modelling procedure of intervention analysis.

An *intervention model* can be conveniently designed for handling more complex situations than that displayed in Figure 19.1.1. Firstly, an intervention model can stochastically model the effects of any number of *interventions* upon the mean level of a series. The external interventions may be man-induced, such as the one in Figure 19.1.1, or caused by a natural event like a forest fire (see the application in Section 19.5.4). Secondly, one or more *missing observations* can be estimated when MLE's are obtained for the parameters in the intervention model (see Sections 19.3 and 19.4). Thirdly, the dynamic influences of one or more *covariate series* upon a single output series can be incorporated into the intervention model (see Sections 19.5 and 22.4). Fourthly, an intervention model can be constructed for handling any combination of the foregoing scenarios. Finally, the *autocorrelated noise* which is not modelled by the multiple interventions and inputs, can be effectively described by an ARMA model.

In a nutshell, intervention analysis is a stochastic modelling technique to analyze rigorously the effects of either man-induced or natural interventions upon the mean level of a time series. The technique of intervention analysis was first suggested by Box and Tiao in 1975 while in the same year, Hipel et al. (1975) introduced the concept into hydrology by ascertaining the effects

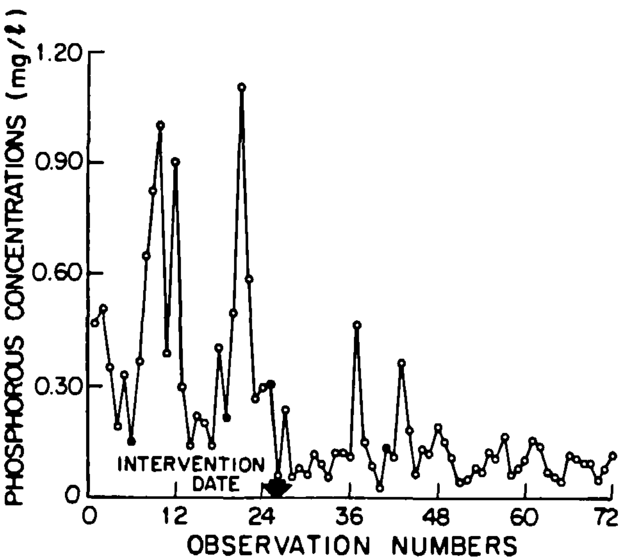


Figure 19.1.1. Monthly phosphorous data (mg/*l*) on the Speed River near Guelph, Ontario, Canada.

of the Aswan Dam upon the mean flows of the Nile River (see Section 19.2.4). As will be seen, the intervention model used in an intervention analysis study is in fact a special type of TFN model and can be used with both seasonal and nonseasonal data. However, due to the great practical importance of intervention analysis, the intervention model is considered in depth in this chapter as well as Chapter 22. The comprehensive design of the intervention model makes it an indispensable tool for use by practitioners in any field where intervention effects must be taken into account. One major area in which the intervention model has been used in the past and will be utilized extensively in the future, is *environmental impact assessment*. As demonstrated by the applications in this book and elsewhere, both natural and man-induced interventions have been modelled for both seasonal and nonseasonal time series in a number of different areas. Below is a list of some of the many fields in which intervention analysis could be quite useful, where the first six categories could be considered to fall within the realm of environmental impact assessment.

Water Quantity: Intervention analysis can be used in hydrology to determine statistically the effects of dam construction on annual (see Section 19.2.4 and also Hipel et al. (1975)) and monthly (see the example given in Section 19.2.5 and also Hipel et al. (1975), other applications are presented in Section 22.4) riverflows. To ascertain the stochastic effects of a forest fire on monthly riverflows, an intervention model is developed in Section 19.5.4 as well as by Hipel et al. (1977b, 1978). Baracos et al. (1981) construct an intervention model to determine whether or not the installation of a new type of snow gauge in the Northwest Territories in Canada introduced a new kind of systematic error into the snow measurements. To determine the impacts of a newly constructed dam on weekly flow rates, Downing et al. (1983) develop an intervention model that includes rainfall inputs. Finally, Shaw and Maidment (1987) employ intervention analysis to ascertain the effects of various water use restrictions upon water demand in the city of

Austin, Texas.

Water Quality: In Section 19.4.5, an intervention model is constructed for the time series charted in Figure 19.1.1. Besides building an intervention for this series, D'Astous and Hipel (1979) also construct an intervention model for assessing the ability of tertiary treatment for reducing the phosphorous levels in a river at another location. In Chapter 22 and also in McLeod et al. (1983), trends are detected and then rigorously modelled using intervention analysis for a wide range of seasonal water quality variables. Additionally, Whitfield and Woods (1984) present interesting case studies where intervention analysis is employed for modelling different kinds of seasonal water quality time records. Moreover, Hipel and McLeod (1989) explain and demonstrate how graphical methods, intervention models, nonparametric trend tests, and regression analysis can be effectively utilized in practice for carrying out intervention and trend assessment studies of water quality time series. Lastly, Zetterqvist (1991) compares three approaches for trend assessment in water quality time series, including a unique approach to intervention modelling.

Air Pollution: Box and Tiao (1975) use intervention analysis to determine if pollution control procedures reduce the average monthly air pollution caused by cars in downtown Los Angeles. Intervention analysis could also be utilized to determine by how much pollution abatement techniques reduce the level of pollutants released by smokestacks into the atmosphere. As is well known, specific kinds of pollutants take part in chemical reactions in the atmosphere which in turn cause acid rain.

Biology: As pointed out by Noakes (1986), in order to manage a biological system, such as a fishery, in an effective manner, decision makers must be able to quantify the impacts of man-induced or natural interventions upon the dynamics of the system. Accordingly, Noakes (1986) employs intervention analysis to model the sharp decline in landing of Dungeness Crab which took place after 1970 along the coast of British Columbia. In another biological systems study, Noakes and Campbell (1992), use intervention analysis for examining yearly shell growth measurements of geoduck clams to indicate changes in the marine environment of Ladysmith Harbour, British Columbia. By applying an appropriate intervention model to an average annual index of standardized geoduck growth for the period from 1907 to 1980, they found that there was a 27% decrease in growth after the initiation of log booming and storage in Ladysmith Harbour starting about 1960. Moreover, an 8% increase in geoduck mean annual growth was coincident with an increase in mean yearly temperature starting in 1920.

Acid Rain: In a trend detection study of acid rain in New York State, Bilonick and Nichols (1983) employ intervention analysis to ascertain whether or not the mean level of depositions of nitrate in precipitation measurements were significantly affected by changes in the method for the analysis for nitrate. The discovery of trend changes in acid rain is studied using exploratory data analysis in Section 22.3.5 of this book and also by McLeod et al. (1983).

Energy: When a nuclear power plant comes into effect, scientists, as well as other concerned groups, may wish to know how the plant alters its environment. One major electrical utility company in the United States took appropriate measurements before and after one of its nuclear plants became operational. By using intervention analysis, the company could determine precisely how the environment was altered.

Business: To determine if governmental controls can reduce the monthly rate of inflation, Box and Tiao (1975) employ intervention analysis. Moreover, Wichern and Jones (1977) utilize intervention analysis to assess the impacts of market disturbances while G. McLeod (1983) uses the technique to investigate the effects of an economic recession on quarterly petrochemical consumption. Finally, to ascertain the impacts of the introduction of directory assistance fees upon the number of requests for telephone numbers, Vandaele (1983, Ch. 14) employs intervention analysis.

Transportation: To determine the effectiveness of seat belt legislation on traffic deaths in Australia, Bhattacharyya and Layton (1979) develop an intervention model. Harvey (1989, Section 7.6) presents a state-space formulation of an intervention model and employs intervention analysis to investigate the consequences of seat belt legislation in the United Kingdom. Another interesting problem would be to examine the influence of raising or reducing fares upon the level of utilization of air transportation.

Other Areas: Because of the numerous kinds of human activity which take place worldwide, it would be possible to produce a very long list of areas where intervention analysis could prove to be very useful. Within the health sciences alone, there could be many potential applications. For example, intervention analysis could be used to see how effective price controls are in controlling cigarette consumption.

As mentioned previously, the main reason for studying a given problem using intervention analysis is to determine the effect of one or more interventions upon the mean level of a series. However, it should be emphasized that intervention analysis is a tool designed for rigorously determining the effects of an intervention upon a given system *after* the intervention comes into play. It is not meant to predict what will happen in the future due to an intervention which has not yet occurred. As a matter of fact, to properly calibrate an intervention model, data are required both before and after the intervention.

To further explain the foregoing point, a practical example is informative. Suppose that in order to reduce acid rain, scrubbers are going to be installed in the smokestacks of chimneys at electrical utilities which use coal. Physically based models from the fields of chemistry, physics and engineering could be used to assist in the design of the scrubbers. Based upon the overall model of the design, the manufacturer may claim that his scrubbers are guaranteed to remove specified levels of different pollutants after installation. Needless to say, this may not be what happens. As is the case with all models, even the physical models which are used in the design of the scrubbers are approximations of how natural processes behave. Furthermore, most engineering designs are usually so complex that it is impossible to accurately model all the components of the design and their interconnections. Consequently, a priori predictions of how a physical system should operate after it is brought into operation can be misleading. What really counts is what actually happens after the intervention of installing the scrubbers takes place. By taking appropriate measurements of pollutant levels both before and after the installation of the pollution abatement equipment, intervention analysis can be used to determine precisely how well the scrubbers work. The best estimate of the actual percentage drop in the mean level of a given pollutant and how much uncertainty or variance is contained in this estimate are the types of information which are of ultimate importance to everyone. Indeed, in environmental disputes which go to court, intervention analysis could prove to be a valuable tool for interpreting how certain pollutants are actually affected by man-induced activities. As shown by the applications in Section 19.5.4 and elsewhere, as information becomes available after the date at which a given intervention took place, the fitted intervention model can be employed for predicting how the intervention will continue in the future to affect the system under consideration.

Prior to the development of intervention analysis, the *Student t distribution* was traditionally used to estimate and test for a change in the average level. However, this procedure is not designed for checking for changes in the mean level of a time series. In a Student *t* test, it is assumed that there is a step change from one mean level to another due to an intervention. Further, the observations before and after the intervention should vary about the two means, normally, independently, and with constant but not necessarily equal variance. These assumptions are almost never satisfied in time series analysis, since a time series is usually autocorrelated, sometimes nonstationary and frequently seasonal. In addition, the change in the mean level of a time series may not take place as a step change.

Besides making statistical statements about the changes in the mean levels of a time series due to one or more interventions, intervention analysis can be utilized for other purposes. Firstly, by using only a few model parameters, the intervention model furnishes an *efficient summary* of the entire data set, including the effects of the intervention. Note that when the intervention analysis is utilized *all* of the observations are used to calibrate the single intervention model. Previously, practitioners would often discard data before or after an intervention since they did not have a single model available to fit to the complete time series. Secondly, in the process of designing an appropriate intervention model to fit to the data and also by the types of parameters included in the final model, the practitioner can gain *insights* into the physical properties of the system being modelled and how it is dynamically affected by the interventions. For a discussion on the physical justification of ARMA models, the reader may wish to refer to Section 3.6. Finally, because an intervention model is a stochastic model, it can be used for other standard purposes like *forecasting and simulation.*

In the upcoming sections of this chapter, important special cases of the general intervention model are introduced until Section 19.5 where the complete intervention model is presented. An intervention model for a single time series acted upon by multiple external interventions is described in Section 19.2. The method for estimating missing data points in a single time series for which there are no interventions is then considered, followed by the presentation of an intervention model for handling situations where there are both missing observations and multiple interventions. Finally, in Section 19.5 the general intervention model is described for modelling a situation where a covariate series is dynamically affected by both multiple interventions and multiple input series, and there are missing observations in the output. The reader who wishes to start by reading about the most general form of the intervention model, may wish to go directly to Section 19.5. For modelling seasonal time series where the correlation structure depends upon the season of the year, a periodic intervention model is presented in Section 19.6. This model is related to the periodic model described in Chapter 14 where a separate AR or ARMA model is developed for each season of the year. Before the conclusions, suggestions are given about how data should be properly collected in order to optimize the ability of intervention analysis to extract information from the collected data.

Throughout this chapter, all of the models are mathematically described and practical environmental applications are used for explaining how intervention models can be easily constructed in practice. In addition to Chapter 19, applications of intervention analysis to both water quantity and quality time series are presented in Section 22.4 of Chapter 22. For the intervention analysis applications in Chapters 19 and 22, the times of the occurrence of the interventions are known. For situations where there may be trends caused by unknown interventions, comments are made in this chapter about how to detect them while extensive explanations regarding the detection of unknown trends are given in Chapter 23 using nonparametric trend tests as well as in Chapter 24 employing regression analysis in combination with graphical displays. Subsequent to detecting the effects of the interventions, an appropriate intervention model can be developed by following the identification, estimation and diagnostic check stages of model construction.

19.2 INTERVENTION MODELS WITH MULTIPLE INTERVENTIONS

19.2.1 Introduction

Often a single time series is influenced by one or more external *interventions*. Consider for example, how the construction of the Aswan dam affected the average annual flows of the Nile River shown in Figure 19.2.1. In 1902, the first dam on the Nile River at Aswan, Egypt, was completed and the reservoir was filled for the first time in 1902-1903. In Figure 19.2.1, average annual values are calculated in $m^3/s{\times}10^3$ for the water year from October 1 to September 30 for each year from October 1, 1870, to September 30, 1945. Notice in the figure, that the man-induced intervention of building a dam appears to have lowered the mean level from 1902 onwards. In fact, the mean of the first 32 average annual values from October 1, 1870, to September 30, 1902, is 3370.12 m^3/s , while from October 1, 1902, to September 30, 1945, the last 43 values have a mean of 2620.41 m^3/s . There is an obvious drop of 749.71 m^3/s or about 22% in the average flow of the Nile River due to the reservoir construction. As shown in Section 19.2.4 for this application, intervention analysis allows for formulating rigorous statistical statements regarding the change in mean flow and also developing a stochastic model that can be used for forecasting and simulation.

In the TFN modelling of Part VII, cause and effect relationships can be easily modelled by incorporating one or more *input series* into the dynamic component of the overall TFN model. For instance, the influence of precipitation upon riverflow could be easily handled by designing an appropriate transfer function which would describe how the precipitation input affects the output of riverflow. Higher or lower precipitation would result in appropriate increases or decreases in the riverflows. However, for the case of the Nile River in Figure 19.2.1, there is no time series available to represent the intervention of dam construction. Consequently, a *dummy series* is constructed to represent quantitatively the occurrence and nonoccurrence of the intervention. This dummy series is referred to as an *intervention series* and is explained in detail in the next section. Based upon an understanding of how the interventions can affect the output, an appropriate transfer function can be designed for describing the effect of the intervention upon the output. Special identification tools are described for deciding upon how the intervention series should be constructed and the parameters which are required in the transfer function used with the intervention series.

Subsequent to designing the parameters required in the entire intervention model, MLE's can be obtained for the model parameters and the model residuals can be subjected to stringent diagnostic testing. As shown by the Nile River application in Section 19.2.4, an automatic selection criterion such as the AIC in [6.3.1] can be quite useful for model discrimination purposes. For the case of the Nile River, the intervention is known in advance. Because the occurrence of interventions may not be known for some applications, the detection of *unknown interventions* is

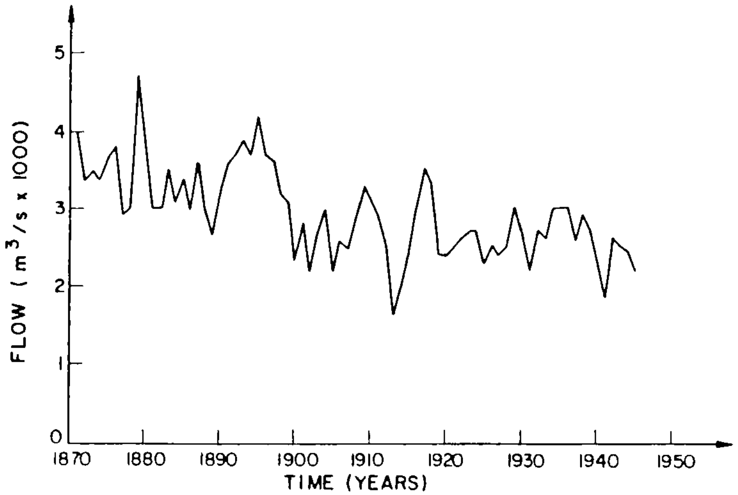


Figure 19.2.1. Average annual flows of the Nile River at Aswan.

discussed in conjunction with model construction in Section 19.2.3 as well as Sections 22.3, 23.3 and 24.2.1. To explain how intervention analysis can be employed with *seasonal data*, the application in Section 19.2.5 is presented where an intervention model is constructed for modelling the stochastic influence of reservoir operation upon average monthly downstream riverflows.

19.2.2 Model Description

Qualitatively, an intervention model with one or more interventions can be written as

response variable = dynamic component + noise

where the dynamic component contains intervention terms for modelling the influences of one or more interventions upon the output or response variable. More precisely, an intervention model with multiple interventions can be described by

$$(y_t - \mu_y) = f(\mathbf{k},\xi,t) + N_t \qquad [19.2.1]$$

where t stands for discrete time, y_t is the response series which may be transformed using a transformation such as the Box-Cox power transformation in [3.4.30], μ_y is the mean of the entire y_t series, N_t is the stochastic noise term which is usually autocorrelated, and $f(\mathbf{k},\xi,t)$ is the dynamic component. The dynamic component includes a set of parameters, **k**, which are needed in the transfer functions and a set of intervention series, ξ, where there is a separate intervention series for each intervention. The dynamic and noise components are now discussed separately.

Dynamic Component

Single Intervention: First consider the situation where there is a single intervention that affects the output y_t. The dynamic component can be written as

$$\begin{aligned} f(\mathbf{k},\xi,t) &= f(\delta,\omega,b,\xi,t) \\ &= \nu(B)\xi_t \\ &= \frac{\omega(B)}{\delta(B)}B^b\xi_t \end{aligned} \qquad [19.2.2]$$

where $\nu(B)$ is the transfer function and ξ_t is the fabricated intervention series. The form of the transfer function is exactly the same as the one described in [17.2.1] for TFN models. In particular, the *transfer function* is given as

$$\begin{aligned} \nu(B) &= \frac{\omega(B)}{\delta(B)}B^b \\ &= \frac{(\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_m B^m)B^b}{(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)} \end{aligned}$$

where $\omega = \{\omega_0,\omega_1,\omega_2,\cdots,\omega_m\}$ is the set of parameters in the operator $\omega(B)$ in the numerator of the transfer function, $\delta = \{\delta_1,\delta_2,\cdots,\delta_r\}$ is the set of parameters in the denominator of the transfer function, b is the *delay time* required for the intervention to affect the output, and $\mathbf{k} = \{\delta,\omega\}$ is the total set of parameters in the transfer function where δ and ω must be estimated from the data. As explained in Section 17.2.2, for stability the roots of the characteristic equation $\delta(B) = 0$ must lie outside the unit circle. The sets of model parameters given by δ and ω are estimated simultaneously with all the model parameters in the complete intervention model in [19.2.1]. In some cases, it may be desirable to calculate the *impulse response weights*, $\nu_0,\nu_1,\nu_2,\cdots$, when the transfer function is written as $\nu(B) = \nu_0 + \nu_1 B + \nu_2 B^2 + \cdots$. Given δ, ω and b, the impulse response weights can be easily calculated using [17.2.2] in the chapter on TFN modelling.

Based upon an understanding of the problem being modelled, the *intervention series*, ξ_t, is designed to consist of a sequence of ones and zeroes where the sequence is the same length as the y_t series. When the intervention is taking place, the series is given a value of one whereas it is assigned a value of zero whenever the intervention is not in effect. Consequently, the intervention series can be thought of as an indicator sequence, since it indicates the presence or absence of the intervention. Two important classes of intervention series which occur quite often in practice are the step and impulse intervention series.

If an intervention takes place as a *step function* at time T, then ξ_t can be represented by the step indicator variable $S_t^{(T)}$ where

$$\begin{aligned} S_t^{(T)} &= 0\,, \quad t < T \\ S_t^{(T)} &= 1\,, \quad t \geq T \end{aligned} \qquad [19.2.3]$$

Figure 19.2.2 shows the step dynamic response given by

$$\frac{\omega(B)}{\delta(B)}B^b S_t^{(T)}$$

which is transferred to y_t for various transfer functions.

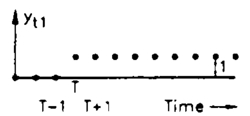
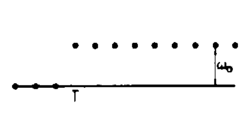
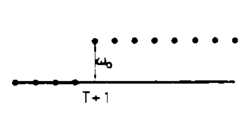
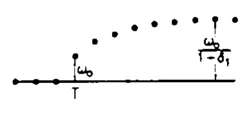
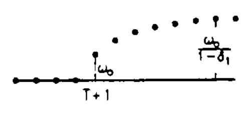
Figure	$\frac{\omega(B)}{\delta(B)} B^b S_t^{(T)}$	Graph of Dynamic Response to a Step Input
a	$S_t^{(T)}$	
b	$\omega_0 S_t^{(T)}$	
c	$\omega_0 B S_t^{(T)}$	
d	$\frac{\omega_0}{1-\delta_1 B} S_t^{(T)}$	
e	$\frac{\omega_0 B}{1-\delta_1 B} S_t^{(T)}$	

Figure 19.2.2. Dynamic response to a step input.

For situations where a step intervention causes an immediate step dynamic response in the output, the model in Figure 19.2.2b may be appropriate. The intervention for the Nile River in Figure 19.2.1 is an example of a step intervention of this type because from 1902 onwards the Aswan dam was operational whereas before 1902 it did not exist. Another example of this kind of step intervention is the construction of a sewage treatment plant that operates continuously after a certain date. This causes a decrease ω_0 in the BOD (biological oxygen demand) level of the receiving body of water. When the step response is not immediate but delayed by time b, then a model of the form shown in Figure 19.2.2c (where b = 1) would be acceptable.

If a step intervention causes a gradual change that asymptotically approaches a limiting step response, then refer to the model in Figure 19.2.2d. The gradual filling of a new reservoir and then the continuous operation of the dam may cause this type of dynamic response in the regulated riverflow patterns. For this case, ω_0 would represent the original change in flow and δ_1, the rate of decay of this change. Intervention models could then be fitted to different periods of the year to indicate, for instance, the change in the new spring and summer flows. When a delay time is also necessary, then the model in Figure 19.2.2e may be the suitable one to use.

The models in Figures 19.2.2d and e (and also Figures 19.2.3d and e) are called *first-order dynamic responses* because the linear difference equations generating these responses are analogous to first-order linear differential equations. For a better interpretation of transfer functions with a term in the denominator, expand the denominator in an infinite series using a Taylor's series. For example, the transfer function in Figure 19.2.2e is

$$\begin{aligned} v(B) &= \frac{\omega_0 B}{1-\delta_1 B} = \omega_0 B(1-\delta_1 B)^{-1} \\ &= \omega_0 B(1+\delta_1 B+\delta_1^2 B^2+\delta_1^3 B^3+\cdots) \\ &= \omega_0 (B+\delta_1 B^2+\delta_1^2 B^3+\delta_1^3 B^4+\cdots) \end{aligned} \qquad [19.2.4]$$

This expanded polynomial then operates on $S_t^{(T)}$ and as shown in Figure 19.2.2e, for a step input $S_t^{(T)}$ the dynamic response increases from time $T+1$ onward (remember delay time is b = 1) to a limiting value $\omega_0/(1-\delta_1)$ which is called the *steady state gain*. Also note that the impulse response weights, $v_0, v_1, v_2, v_3, \ldots,$ can be obtained directly from [19.2.4] by comparing coefficients of B^k, $k=0,1,2,\ldots,$ in $v(B) = v_0 + v_1 B + v_2 B^2 + v_3 B^3 + \cdots$, to those in [19.2.4]. Consequently, $v_0 = 0$, $v_1 = \omega_0$, $v_2 = \omega_0\delta_1$, $v_3 = \omega_0\delta_1^2$, and in general $v_k = \omega_0\delta_1^{k-1}$. Because $|\delta_0| < 1$ for a stable system, the impulse response function decreases for increasing lag k to a limiting value of zero. After determining the impulse response weights, the aforementioned steady state gain is calculated from the definition in [17.2.3] to give a value of

$$g = 0 + \omega_0 + \omega_0\delta_1 + \omega_0\delta_1^2 + \cdots = \frac{\omega_0}{1-\delta_1}$$

for the transfer function in Figure 19.2.2e.

If an intervention takes place as a *pulse input* at time T, then ξ_t can be portrayed by the pulse indicator variable $P_t^{(T)}$, where

$$\begin{aligned} &P_t^{(T)} = 0, \quad t \neq T \\ &P_t^{(T)} = 1, \quad t = T \end{aligned} \qquad [19.2.5]$$

Figure 19.2.3 shows the pulse dynamic responses for different transfer functions. It should be noted that since

$$(1-B)S_t^{(T)} = P_t^{(T)}$$

then it is possible to change all the pulse responses in Figure 19.2.3 to step responses in Figure 19.2.2 by multiplying $P_t^{(T)}$ by $(1-B)^{-1}$.

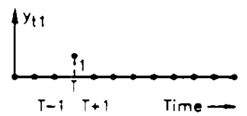
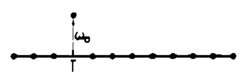
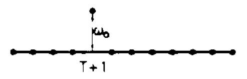
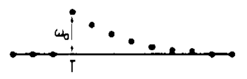
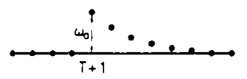
Figure	$\frac{\omega(B)}{\delta(B)} B^b P_t^{(T)}$	Graph of Dynamic Response to a Pulse Input
a	$P_t^{(T)}$	
b	$\omega_0 P_t^{(T)}$	
c	$\omega_0 B P_t^{(T)}$	
d	$\frac{\omega_0}{1-\delta_1 B} P_t^{(T)}$	
e	$\frac{\omega_0 B}{1-\delta_1 B} P_t^{(T)}$	

Figure 19.2.3. Dynamic response to a pulse input.

Pulse interventions often occur in water resources and environmental engineering. For example, a certain chemical process at a water treatment plant may be introduced on a trial basis for one day to see if it significantly affects the quality of the water that is then distributed to the consumers. If the effects of this treatment are delayed one day due to distribution and storage time, then Figure 19.2.3c may be the correct model. Here, ω_0 would represent the water quality change being measured.

The felling of a large number of trees for lumber in a small river basin may act as a pulse intervention and affect the riverflow so that the first-order model in Figure 19.2.3d may adequately describe the resulting change in riverflow. In this model, ω_0 would indicate the initial change in flow, and δ_1 the rate of decay of the change as new trees mature over the years. An intervention term similar to this is developed in Section 19.5.4 for describing the impacts of a forest fire upon riverflows.

Multiple Interventions: By introducing an additional subscript, the intervention component in [19.2.2] can be extended for handling any number of external interventions. If there are I_1 interventions acting upon a single series, y_t, the dynamic component of the intervention model is

$$f(\mathbf{k},\xi,t) = f(\delta,\omega,\mathbf{b},\xi,t)$$
$$= \sum_{i=1}^{I_1} v_i(B)\xi_{ti} \qquad [19.2.6]$$

where ξ_{ti} is the ith fabricated intervention series consisting of 1's and 0's to indicate the presence and absence of the ith intervention, respectively; $\mathbf{k} = (\delta,\omega,\mathbf{b})$ is the set of model parameters where the δ and ω parameters are usually estimated from the data and $\mathbf{b} = \{b_1,b_2,\cdots,b_{I_1}\}$ is the set delay times for the interventions to affect the output. The ith transfer function, which reflects the manner in which the ith intervention affects the output, is written in the same manner as in [17.5.2] for a TFN model as

$$v_i(B) = \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}$$
$$= \frac{(\omega_{0i} - \omega_{1i}B - \omega_{2i}B^2 - \cdots - \omega_{m_i i}B^{m_i})B^{b_i}}{(1 - \delta_{1i}B - \delta_{2i}B^2 - \cdots - \delta_{r_i}B^{r_i})}$$

where m_i and r_i are the orders of the operators $\omega_i(B)$ and $\delta_i(B)$, respectively; and b_i is the delay, specified as a positive integer, before the ith intervention affects y_t. Notice that the ith transfer function in [19.2.6] is identical to the one in [19.2.2] except that the subscript i has been added to indicate that $v_i(B)$ is the transfer function for the ith intervention series, ξ_{ti}.

As illustrated by the applications in this chapter and also Section 22.4, usually only a few parameters are required in each transfer function and therefore, m_i and r_i are 0 or 1. After estimating the parameters in the $\omega_i(B)$ and $\delta_i(B)$ operators along with all the other parameters in the complete intervention model, it may be required to calculate the impulse response weights v_{ji}, j=0,1,2, . . . , in the operator

$$v_i(B) = v_{0i} + v_{1i}B + v_{2i}B^2 + \cdots$$
$$= \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}$$

This can be easily accomplished by following the procedure outlined in Section 17.2.2.

Noise Term

After modelling the effects of the interventions upon the output, the noise term describes what cannot be modelled by the dynamic component as

$$N_t = y_t - f(\mathbf{k},\xi,t)$$

As is the case for the TFN models of Chapters 17 and 18, usually the noise term can be effectively explained by the ARMA model in [3.4.4], [16.2.3], [16.2.4] or [17.2.4]. Consequently, a model for the noise is

$$\phi(B)N_t = \theta(B)a_t$$

or

$$N_t = \frac{\theta(B)}{\phi(B)}a_t \qquad [19.2.7]$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA operators of order p and q, respectively, and a_t is the white noise which is $NID(0,\sigma_a^2)$. When differencing is required to remove nonstationarity, the noise term, N_t, can be modelled using the ARIMA model in [4.3.4].

Complete Intervention Model

To simultaneously model both the effects of one or more interventions upon the output and the remaining correlated noise contained in the system, the dynamic and noise components can be combined to form the intervention model. For the situation where there is a single intervention, the intervention model is formulated using [19.2.2] and [19.2.7] as

$$\begin{aligned} y_t - \mu_y &= v(B)\xi_t + N_t \\ &= \frac{\omega(B)}{\delta(B)}B^b\xi_t + \frac{\theta(B)}{\phi(B)}a_t \end{aligned} \qquad [19.2.8]$$

When there are I_1 external interventions which influence y_t, the overall intervention model is derived using [19.2.6] and [19.2.7] to be

$$\begin{aligned} y_t - \mu_y &= \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + N_t \\ &= \sum_{i=1}^{I_1} \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}\xi_{ti} + \frac{\theta(B)}{\phi(B)}a_t \end{aligned} \qquad [19.2.9]$$

Effects of an Intervention Upon the Mean Level

As indicated earlier, one of the main purposes of intervention analysis is to ascertain the change in the mean level of a series due to one or more interventions. Because the impacts of a given intervention upon the output y_t are reflected by the magnitude of the parameters in the transfer function, it would be expected that the change in the mean level is a function of the transfer function parameters. To calculate the change in the mean level, first determine the expected value of y_t before the intervention to obtain $E[y_t]_{before}$ and then ascertain the expected value of y_t after the intervention to get $E[y_t]_{after}$. The change in the mean level is then simply determined using

$$change = E[y_t]_{after} - E[y_t]_{before} \qquad [19.2.10]$$

When the percentage change in the mean level of y_t due to the intervention is required, it can be calculated using

$$\% \, change = \left(\frac{E[y_t]_{after} - E[y_t]_{before}}{E[y_t]_{before}} \right) 100 \qquad [19.2.11]$$

If the original series were transformed using the Box-Cox transformation in [3.4.30], in order to obtain the mean level change in terms of the untransformed series, the inverse Box-Cox transformation must be determined before calculating the expected values and substituting them into [19.2.10] or [19.2.11].

Example with a Step Intervention: Consider the case for [19.2.8] where there is a single step intervention as in [19.2.3] which takes place at time t = T and ω_0 is the parameter in the transfer function. Hence, the intervention model is written as

$$y_t - \mu_y = \omega_0 \xi_t + N_t \qquad [19.2.12]$$

where

$$\xi_t = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}$$

and μ_y stands for the mean level of the entire response series. Because the noise term is assumed to be the same before and after the intervention, the exact form of the noise term does not matter when calculating the change or percentage change in the mean level. Before the intervention $\xi_t = 0$ and, therefore,

$$y_t - \mu_y = N_t \quad \text{for } t < T$$

Taking expected values

$$E[y_t]_{before} = E[\mu_y] + E[N_t]$$

Because the expected value of a constant is itself

$$E[y_t]_{before} = \mu_y + \frac{\theta(B)}{\phi(B)} E[a_t]$$

But $E[a_t] = 0$ and consequently the above simplifies to

$$E[y_t]_{before} = \mu_y \qquad [19.2.13]$$

After the intervention, $\xi_t = 1$ and hence the intervention model is

$$y_t - \mu_y = \omega_0 + N_t \quad \text{for } t \geq T$$

Upon taking expected values of the above

$$E[y_t]_{after} = \mu_y + \omega_0 + \frac{\theta(B)}{\phi(B)} E[a_t]$$

$$= \mu_y + \omega_0 \qquad [19.2.14]$$

The change in the mean level is calculated using [19.2.10] to be

$$change = ((\mu_y + \omega_0) - \mu_y) = \omega_0 \qquad [19.2.15]$$

Utilizing [19.2.11], the percentage change is

$$\% \ change = \left(\frac{(\mu_y + \omega_0) - \mu_y}{\mu_y} \right) 100$$

$$= \frac{\omega_0}{\mu_y} 100 \qquad [19.2.16]$$

Example with a Logarithmic Data Transformation and a Step Intervention: Suppose that the intervention model is the same as in the first example except for the fact that the data were first transformed using natural logarithms. Equations [19.2.10] and [19.2.11] could be utilized to obtain the change and percentage changes in the mean levels for the logarithmic data. However, to determine the mean changes in the original untransformed series represented by Y_t, take anti-logarithms of $y_t - \mu_y = \omega_0 \xi_t + N_t$ to obtain

$$Y_t = exp(\mu_y + \omega_0 \xi_t + N_t)$$

$$= e^{\mu_y} e^{\omega_0 \xi_t} e^{N_t} \qquad [19.2.17]$$

Before time t = T, each value of ξ_t is zero and hence

$$Y_t = e^{\mu_y} e^0 e^{N_t} = e^{\mu_y} e^{N'}$$

Taking expected values gives

$$E[Y_t]_{before} = E[e^{\mu_y} e^{N_t}]$$

$$= e^{\mu_y} E[e^{N_t}] \qquad [19.2.18]$$

where e^{μ_y} is a constant. After the intervention, ξ_t possesses a value of unity and, therefore,

$$y_t = e^{\mu_y} e^{\omega_0} e^{N_t}$$

By taking expected values,

$$E[Y_t]_{after} = E[e^{\mu_y} e^{\omega_0} e^{N_t}]$$

$$= e^{\mu_y} e^{\omega_0} E[e^{N_t}] \qquad [19.2.19]$$

since e^{μ_y} and e^{ω_0} are constants. An advantage of calculating the percentage change in the mean level is the factor $E[e^{N_t}]$ drops out of the expression and therefore does not have to be estimated. Hence, using [19.2.11], the percentage change in the mean is

$$\% \ change = \left(\frac{e^{\mu_y} e^{\omega_0} E[e^{N_t}] - e^{\mu_y} E[e^{N_t}]}{e^{\mu_y} E[e^{N_t}]} \right) 100$$

$$= (e^{\omega_0} - 1)100 \qquad [19.2.20]$$

When a confidence interval is required for the percentage change, this can easily be calculated using the above equation. Suppose, for instance, the 95% confidence interval were needed. Because the MLE for ω_0 is approximately normally distributed, then $\hat{\omega}_0 \pm 1.96SE$ could be substituted into the above equation. Hence, the upper limit would be

$$(e^{\hat{\omega}_0 + 1.96SE} - 1)100$$

and the lower limit would be

$$(e^{\hat{\omega}_0 - 1.96SE} - 1)100$$

where the best estimate of the percentage change is

$$(e^{\hat{\omega}_0} - 1)100.$$

19.2.3 Model Construction

In many situations, the fact that one or more interventions has taken place is known and the analyst wishes to design an intervention model to describe changes which may have occurred in the output. For example, when a pollution abatement procedure is implemented, an intervention model can be constructed for ascertaining how effective the procedure is for reducing the level of the pollutant. In Section 19.4.5, an intervention model is developed for statistically determining how much the phosphorous levels in the Speed River shown in Figure 19.1.1, have been reduced by tertiary sewage treatment. For describing the effects of reservoir construction upon the average annual flows of the Nile River displayed in Figure 19.2.1, an appropriate intervention model is constructed in Section 19.2.4. Other time series which have been influenced by known interventions, are modelled using intervention analysis in upcoming sections of this chapter as well as in Section 22.4.

In some instances, *unknown interventions* may cause unexpected trends to occur in the data. For example, if measuring equipment becomes faulty due to over usage, the scientist may not be initially aware that a systematic measuring error has been introduced into his data. An owner of a factory may illegally dump liquid wastes into a receiving body of water in order to avoid paying for the treatment of his wastes. Environmentalists who monitor the affected stream would certainly like to detect and model the affects of the initially unknown industrial pollution. The graphical techniques of Sections 22.3 and 24.2.2 as well as the nonparametric trend tests of Chapter 23 can be used for detecting trends in water quality and other kinds of time series, which may be caused by unknown or suspected interventions.

Even if at least one intervention is known to have occurred, other unknown interventions may create unsuspected trends in the time series which is being studied. Consequently, as shown in Figure 19.2.4, prior to constructing an intervention model by following the usual three stages of model construction discussed in previous chapters, it is recommended that simple detection procedures be implemented for discovering statistical anomalies which may be caused by unknown interventions. This is especially true when one is dealing with the type of *messy environmental data* studied in Part X, where the data collection schemes may not have been carefully designed and land use changes, which may have been known when they were initiated, were not properly recorded. When the reasons for the unknown trends have been accounted for, an appropriate intervention model can be developed by following the remaining steps in Figure 19.2.4. Based upon a knowledge of the interventions which were previously known and also those which were discovered at the detection stage, an intervention model can be designed for describing what is expected to occur. To quantify what is hypothesized to take place, appropriate intervention series and accompanying transfer functions must be decided upon. Additionally, a tentative noise model must be selected. Following this, the parameters of the noise model and transfer functions are estimated using the method of maximum likelihood. Then the model is checked for possible inadequacies. Problems with the model residuals, for example, may indicate trends caused by an intervention which was not found at the detection stage. If discrepancies are observed, then suitable model modifications can be made. The construction of an intervention model is now discussed, with special emphasis being placed on the detection of trends and identification of an intervention model to describe the trends.

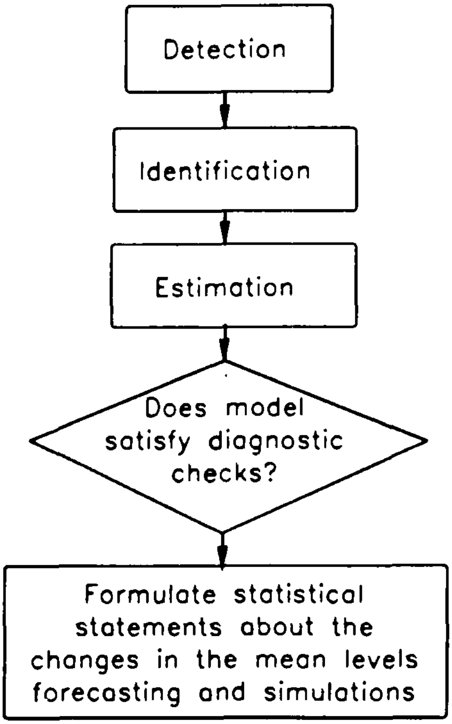

Detection
Identification
Estimation
Does model satisfy diagnostic checks?
Formulate statistical statements about the changes in the mean levels forecasting and simulations

Figure 19.2.4. Constructing an intervention model.

Detection

Exploratory Data Analysis: In order to detect trends in a time series which may be caused by unknown interventions, simple statistical procedures can be used. Employing these straightforward yet informative statistical methods for the detection of trends, can be considered as part of the statistical methodology which Tukey (1977) calls exploratory data analysis. As pointed out in Section 1.2.4, the objective of exploratory data analysis is to uncover important statistical characteristics of the data, such as the presence of various kinds of trends, by carrying out numerical and graphical detective work. Usually, graphs of various statistics constitute the most effective and convenient approach for interpreting how a given time series generally behaves and the overall manner in which trends may occur due to both unknown and known interventions.

For the intervention analysis applications considered in this chapter, the exact times when all of the interventions began are known. However, in Chapter 22, where a wide variety of water quality series are examined, in some cases the times when possible interventions started are not known a priori. Consequently, a detailed explanation of useful exploratory data analysis tools which can be used for detecting trends caused by unknown interventions, is presented in Section 22.3 of Chapter 22 rather than in this section. The specific exploratory data analysis tools which are discussed include:

1. plots of the time series;
2. box-and-whisker graphs (Tukey, 1977);
3. cross-correlations;
4. Tukey smoothing (Tukey, 1977; Velleman and Hoaglin, 1981);
5. autocorrelation function.

Practical applications are utilized in Chapter 22 for demonstrating the efficacy of the foregoing methods for discovering important statistical properties of different kinds of water quality time series. Moreover, the trend analysis studies of water quality time series measured in rivers which are presented in Section 24.3, illustrate how the robust locally weighted regression smooth of Cleveland (1979) outlined in Section 24.2.2 can be employed for visually detecting trends.

The authors wish to emphasize that even when it is known in advance that certain interventions have occurred during known time periods, it is usually advisable to still employ relevant detection tools for discovering the effects of unknown interventions and better understanding how both known and newly discovered interventions have influenced the behaviour of the series. This is especially true when it is suspected that reliable personnel and/or equipment were not used for collecting specific data and recording events that could cause trends in the data. Whatever the case, after one or more unexpected trends are detected using exploratory data analysis, appropriate historical documentation should be searched to see if a physical reason can be found. For example, a suspected pollution spill that may have occurred in a river may not be recorded by the agency that collected the water quality data but it may be written down by another institution which is concerned with enforcing water quality standards. Only if a reasonable physical reason can be found for explaining the presence of an unexpected trend should intervention analysis be used to rigorously ascertain the effects of the intervention at the confirmatory data analysis stage. In some cases, what is thought to be a trend due to some external physical cause may in fact only be a stochastic trend which operates according to probabilistic laws. As explained in Section 4.6, the stochastic trend may be suitably described by a stochastic model which does not have an intervention component. The reader should keep in mind that even when simulating an autocorrelated sequence with a stationary model, there can be relatively long periods of time during which the level of the series remains either entirely above or below the mean level (see Figures 2.3.2 and 2.3.3). Furthermore, even though the probability of occurrence is low, some sequences of synthetic data may continually increase or decrease over certain time periods and, therefore, may appear to be deterministic trends. Consequently, when a thorough investigation of a given series indicates that a certain trend is not caused by an external intervention, then it should be properly modelled as a stochastic trend.

Other Trend Detection Techniques: In addition to the simple exploratory data analysis tools, some of which are thoroughly discussed in this book in Chapter 22, other methods are available for detecting trends. As reported by MacNeill (1980), the problem of testing for changes in the parameters of a regression model at an unknown times was first investigated by Quandt (1958, 1960) who developed a likelihood ratio test for no change versus one change. Further research by Hinkley (1969) and Feder (1975) also dealt with the likelihood ratio test approach. Brown et al. (1975) suggested tests based upon recursively generated residuals and the associated sequence of partial sums of these residuals. Following this, MacNeill (1978a,b) investigated the properties of sequences of partial sums of raw regression residuals and proposed a Cramer-von Mises type of statistic for testing for change of regression at an unknown time. As an alternative approach to his earlier work, MacNeill (1980) proposed a new method based on a likelihood ratio type of test for discovering changes in regression when the change times are unknown. The test statistic of MacNeill (1980) was derived utilizing an approach of Chernoff and Zacks (1964), Gardner (1969) and MacNeill (1974) for detecting parameter changes at unknown times when the random variables are IID. To demonstrate the usefulness of his approach, MacNeill (1980) applied his test to various climatological data sets. Additionally, MacNeill (1985) expanded his research published in 1980 and gave further details about a *change-detection statistic* for discovering parameter changes in a time series which occur at unknown times. The overall procedure, referred to by MacNeill (1985) as the adaptive forecasting and estimation using change-detection, was applied to the average annual flows of the Nile River at Aswan (see Section 19.2.4 for an intervention analysis study of this data). More recently, Jandhyala and MacNeill (1989, 1991) as well as Tang and MacNeill (1993) have extended research on the change-point statistic. Finally, MacNeill et al. (1991) have applied the change-point statistic and other trend detection methods to the average annual flows of the Nile River shown in Figure 19.2.1.

Bagshaw and Johnson (1977) proposed procedures for sequentially monitoring forecast errors in order to detect changes in a time series model. Their methods are founded upon likelihood ratio statistics consisting of cumulative sums. To test for changes in the parameter values of an ARIMA model, Bagshaw and Johnson (1977) extended the work of Page (1954, 1955) which dealt only with mean changes in forecast errors.

Additional procedures for detecting and modelling changes in a process are discussed in Section 24.2.1. Moreover, a range of other useful change detection methods can be found in the literature. For example, Wichern et al. (1976) devise a two-stage method for finding step changes of variance for the case of an AR(1) model. Using a generalized likelihood ratio, Fiorina and Maffezzoni (1979) develop a direct approach to jump detection in linear time-invariant systems. Brillinger (1989) presents a trend test for finding a monotonic trend in a time series. Finally, Kenett and Zacks (1992) propose a new class of tracking algorithms for processes which change their stochastic structure at unknown epochs.

All of the foregoing techniques discussed in the last three paragraphs for detecting unknown changes assume that a formal model is first fitted to the data in order to employ a given test statistic which may be fairly complicated to use in practice. On the other hand, for the simple *graphical exploratory tools* discussed in Section 22.3, no underlying model is assumed. Instead, the given data are visually studied using only simple graphical procedures that can assist the practitioner in detecting the obvious statistical traits such as trends caused by unknown interventions, in addition to other general statistical characteristics. Subsequent to using exploratory data analysis tools, some people may wish to use more formal procedures for detecting unknown interventions to see if they agree with what is found from more qualitative graphical inspections. For instance, the *nonparametric trend tests* of Section 23.3 can be employed for detecting trends in a data set prior to fitting a more sophisticated parametric model such as the intervention model of this chapter. However, in all cases practitioners are advised to first use simple detection tools before employing more formal procedures. Sometimes obvious anomalies in a time series can be missed because the modeller becomes too involved with the technical details of using sophisticated testing procedures.

Within this text, exploratory data analysis tools are employed for gathering information that is eventually used in the design of an appropriate intervention model. If for some reason an unknown intervention is not detected prior to fitting a formal model, anomalies in the residuals of the fitted model may reveal the presence of the impacts of the undetected intervention. Based upon this and other information, a proper intervention model can be designed to realistically account for the impacts due to all the interventions.

Identification

After a practitioner is satisfied that he or she has detected all the possible trends in the data and found reasonable physical explanations as to what may have caused them, he or she can proceed to design an intervention model to formally model the series. As revealed in Figure 19.2.4, the general model construction stages subsequent to the detection phase, are similar to those advocated for use with other time series models such as the nonseasonal model building methods of Part III. In addition to a thorough understanding of the problem plus information uncovered at the detection stage, identification procedures can be used to ascertain which parameters to include in the intervention model in [19.2.9]. This involves designing an intervention series and corresponding transfer function to account for the stochastic effects of each intervention upon the output, and also selecting a tentative noise model. Some of the identification methods in this section could perhaps be considered as exploratory data analysis techniques. However, since they are used mainly for deciding upon which parameters to include in the model, they are described in this section. Because the three stages of model construction after the detection stage are used for developing the most appropriate model to formally model the data, these three stages are in fact part of what Tukey (1977) calls confirmatory data analysis. The fitted intervention model is used to rigorously confirm in a mathematical sense how the interventions have statistically affected the mean level of the series. In other words, quantitative measures of the statistical effects of the interventions are obtained by fitting an intervention model to the data. The exploratory data analysis results really only provide qualitative interpretations of what may be happening. General and specific discussions of data analysis are presented in Section 1.2.4 and Chapter 22, respectively.

In essence, identification permits a qualitative understanding of a given intervention problem that allows it to be converted into a form which can be quantified. This is affected by identifying the appropriate parameters to include in the model in order to check the practitioner's hypothesis about how he thinks the system was affected by one or more interventions. The parameters required in the dynamic and noise components are decided upon separately.

Designing the Dynamic Component: For the case of the model in [19.2.6] or [19.2.9], the only terms in the dynamic component are those which model the impacts of the interventions. The two basic steps to identify the intervention or dynamic component are to:

(1) Ascertain the type of changes in the time series due to the interventions. In other words, use appropriate information to make hypotheses about how the series has been influenced by the interventions.

(2) For each intervention, select an appropriate intervention series and associated transfer function to permit quantification of how the intervention has affected the series.

As noted in Section 19.2.2, an intervention series is a fabricated sequence which is designed to indicate the occurrence and non-occurrence of the interventions. When the intervention is taking place, an entry in the intervention series is assigned a value of 1 while it is given a magnitude of 0 when the intervention is not occurring. Two important classes of intervention series are the step and pulse intervention series given in [19.2.3] and [19.2.5], as well as Figures 19.2.2a and 19.2.3a, respectively. The transfer function for a given intervention series must be selected in such a way that the geometric shape of the dynamic response mimics the geometrical pattern of the trend caused by the intervention in the actual series . For the cases of the step and pulse interventions, the shapes of various dynamic responses are illustrated in Figures 19.2.2 and 19.2.3, respectively. When modelling seasonal data, if the intervention affects certain seasons in a particular manner, an intervention term, consisting of an intervention series and associated transfer function, can be designed for each season or group of seasons that are changed in the same fashion. This point is clarified by the intervention models developed for seasonal data in Sections 19.2.5, 19.4.5 and 19.5.4.

Various techniques are available to use in step 1. For nonseasonal data, a plot of the time series should reveal how the series differs before and after each intervention. If the observations are seasonal, then in addition to a plot of the series, one or more of the graphical methods shown presently may prove useful. These different approaches are described for the general case when there are s seasons per year. For specific types of seasonal data, such as quarterly and monthly data, s is simply assigned the correct values, like 4 and 12, respectively. For each method, every season over all the years is analyzed to see how each intervention affected that season. Nonseasonal data can also be analyzed by the following methods. Also note that some of the information described here may already be available from graphical studies executed at the detection stage.

(1a) *Seasonal plots.* A graphical display for each individual season over all the years on record should reveal specific seasons that are affected by the intervention and in what manner they have changed. Keeping in mind that the seasonal plots contain the dynamic component plus the noise term, transfer functions and intervention series can be designed to obtain dynamic responses that model the seasonal interventions. If it is thought that the response variable may require a transformation such as natural logarithms, then seasonal plots may be made of the transformed data.

(1b) *Cusum chart.* The cumulative sum (cusum) technique was proposed by Page (1954) and Barnard (1959) and improved upon by Lucas (1985) and others. The cumulative sum is calculated and then plotted for each season to see how the seasonal average changes after the intervention. Let the data for season i over N years be denoted by $y_{1i}, y_{2i}, \cdots y_{Ni}$. Define the kth cusum CS_{ki} for season i as:

$$CS_{ki} = CS_{k-1,i} + (y_{ki} - \bar{y}_{bi}) = \sum_{j=1}^{k} y_{ji} - k\bar{y}_{bi}, \quad k = 1,2, \cdots ,N \qquad [19.2.21]$$

where $CS_{0i} = 0$ and $\bar{y}_{bi}$ is mean of season i before the intervention.

A cusum chart is a plot of the cusum against time. Before the start of the intervention, the cusum should follow a horizontal line with values fluctuating around that line. However, if after the intervention there is a step intervention and the mean increases to a new level, the cusum will follow a constant upward slope as shown in Figure 19.2.5. If the average for a particular season decreases a constant amount, then after the intervention the cusum will follow a fixed downward slope as illustrated in Figure 19.2.6. The steeper the slope the greater is the step increase or decrease in the average for a particular month. As stated by Woodward and Goldsmith (1964), one of the main advantages of the cusum technique is its sensitivity. Relatively small changes in the mean value appear as distinctly different slopes.

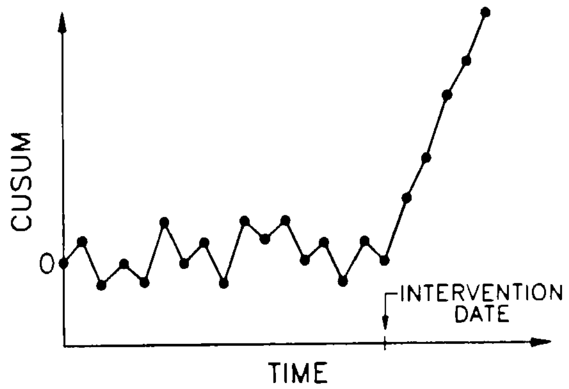


Figure 19.2.5. Cusum chart for a step increase in mean.

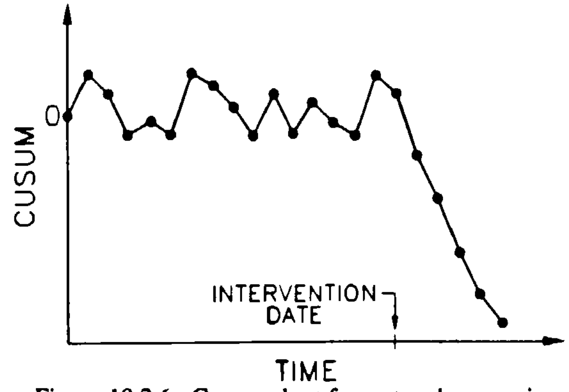


Figure 19.2.6. Cusum chart for a step decrease in mean.

When the mean level for a season increases gradually to a new level, this will be reflected in the cusum chart by a slowly changing slope after the start of the intervention to a steeper constant slope when the mean reaches its new level. This type of average change is illustrated in Figure 19.2.7.

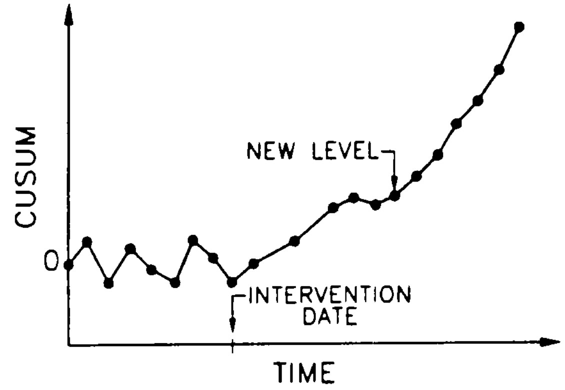


Figure 19.2.7. Cusum chart for a gradual increase in mean to a new level.

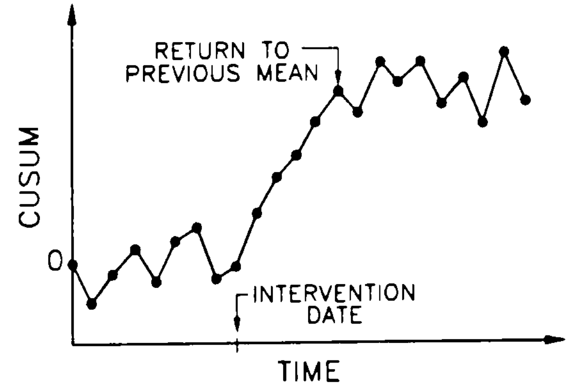


Figure 19.2.8. Cusum chart for a step increase in the mean followed by a step decrease to the previous average.

In general, it is necessary to study a particular cusum chart individually to determine how the mean has been affected by the intervention. If for example, there is a step increase in the mean level due to an intervention and then later the level returns to its former mean, the cusum results for this case are shown in Figure 19.2.8. Notice that a step return to the mean prior to the intervention is reflected in the cusum chart by the cusum once again following a horizontal line. However, the new horizontal line is at a different level than the one before the intervention.

(1c) *Average plots.* Calculate the s seasonal means for all the years up until the intervention. From the intervention onwards calculate the seasonal means for each year after the intervention until the end of the data or start of a new intervention. A useful graph to then plot is the s seasonal averages before the intervention. For each year after the intervention, plot the new seasonal averages on the same graph as the averages before the intervention. Appropriate interpretations concerning the intervention impacts can be drawn by observing how the seasonal averages are affected each year after the intervention.

(1d) *Other plots*. For a particular problem the researcher can of course develop any appropriate aids for model identification to use in conjunction with practical engineering judgement. However, he should keep in mind that for seasonal data, it is often most informative to plot each season separately. As explained and illustrated in Section 22.3.3, for example, one can plot box-and-whisker graphs (Tukey, 1977) for each season both before and after an intervention.

Designing the Noise Term: One or both of the following approaches may be useful to identify the parameters required in the ARMA model for N_t in [19.2.7] and [19.2.9]. The first procedure uses the data before the intervention while the second method utilizes all of the available information.

(1) Following the procedures of Chapters 5 to 7, identify an ARMA model for the response series, y_t, up to the time of the first intervention. Of course, this method can only be used if sufficient data are available before the time of the first intervention. Hence, there should be at least 40 or 50 observations before the intervention. For the special case where there is a single step intervention where the dynamic response is modelled as $\omega_0\xi_t$ as in Figure 19.2.2b, the data after the intervention can be used to identify the form of the ARMA noise term. In general, for any interval of the time series for which the effects of one or more interventions can be neglected or somehow removed, that portion of the data can be used for identifying the form of N_t.

(2) The second technique is the same as the empirical identification procedure in Sections 17.3.1 and 17.5.3 for deciding upon the form of N_t in a TFN model possessing one or more covariate series. After identifying the form of all the intervention terms in the dynamic component, fit the model in [19.2.9] to the series where it is assumed that the noise term is white and hence the intervention model has the form

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + a_t \qquad [19.2.22]$$

In practical applications, usually the noise term is correlated. Consequently, after obtaining the estimated residual series, a_t, for the above model using the method of maximum likelihood, the type of ARMA model to fit to the noise series can be determined by following the three stages of model development described in Chapters 5 to 7. By using the identified form of N_t for the noise term along with the previously designed dynamic component, the intervention model in [19.2.9] is now completely designed.

Estimation

At the estimation stage, MLE's and corresponding standard errors can be simultaneously obtained for all the model parameters in [19.2.9]. In addition, the estimated residual series, $\hat{a}_t$, can also be obtained for use in diagnostic checking. Because an intervention model is simply a specific type of TFN model, the estimation procedure for TFN models, which is mentioned in Section 17.3.2 and described in detail in Appendix A17.1, can be used. In addition, automatic selection criteria such as the AIC in [6.3.1] and the BIC in [6.3.5] can be employed to assist in selecting the most appropriate model. The reader can refer to Figure 6.3.1 for an outline of how an automatic selection criterion such as the AIC can be incorporated into the three stages of model construction.

Box and Tiao (1975) show how the transfer function parameter estimates depend on the y_t series plus the other parameters in the intervention model. These estimates can be shown to be a function of the difference between a weighted average of the y_t's before and after the intervention.

Diagnostic Checking

All of the residual diagnostic checks given in Chapter 7 can be used for verifying the suitability of the fitted intervention model. As noted before, for checking that the residuals are white the recommended procedure is to plot the RACF (residual autocorrelation function) in [7.3.1] along with the 95% confidence limits. In addition, the cumulative periodogram in [2.6.2] and the modified Pormanteau test in [17.3.8] can be used to ascertain whether or not the residuals are uncorrelated. If the residuals are correlated, this implies that the model is inadequate and a more appropriate model can be found by repeating the earlier stages of model construction in Figure 19.2.4. When the residuals are not approximately normally distributed and/or are heteroscedastic, an appropriate Box-Cox transformation of the y_t series using [3.4.30] may rectify the situation.

19.2.4 Effects of the Aswan Dam on the Average Annual Flows of the Nile River

Case Study Description

Within this section and the next one, practical applications are used for demonstrating how intervention models can be conveniently constructed for modelling both nonseasonal and seasonal time series, respectively, which have been affected by external interventions. For the case of the average annual flows of the Nile River at Aswan, the affect of the completion of the Aswan dam in 1902 upon the riverflows are graphically illustrated in Figure 19.2.1. As pointed out in Section 19.2.1, from 1902 onwards, there appears to be a significant drop in the mean level of the flows.

The average flows of the Nile River plotted in Figure 19.2.1, are obtained from a report by Hurst et al. (1946, p. 125). Prior to 1903, levels on the Nile River were measured downstream from the dam site. However, from 1903-1939, discharges were determined accurately by relating sluice measurements of the dam to the downstream gage stages. The rating curve obtained in the period 1903-1939 was used to determine the discharges before 1903. From 1903 to 1945 the discharges are the actual sluice measurements.

The dam intervention that caused a drop in the average flow of the Nile could be an accumulative effect of the following factors (Hurst et al., 1946; Yevjevich and Jeng, 1969).

1. The reservoir size allowed for evaporation losses, greater percolation into the underlying soil, plus other natural losses.
2. Water was taken from the reservoir to be used for irrigation, domestic water supply, and other human-oriented uses.

3. Systematic errors were introduced into the data prior to 1903 by using a rating curve developed from 1903 to 1939. During construction of the dam, channels downstream were opened through the cataracts with a consequent change in the distribution of velocity across the section. This may have caused a change in the gage-discharge relationship. These measurement errors are thought not to exceed 5% (Hurst et al., 1946, p. 23).

Notice in Figure 19.2.1 that the annual flows from October 1, 1899, to October 1, 1902, have values closer to those in the period from 1903 onward when the dam was operating. It could be that the starting of construction of the dam and channel improvements should be considered as the start of the intervention. However, for this analysis the start of the dam operation and reservoir filling in 1903 is considered as the date of intervention. If 1899 were considered as the intervention date, parameter values for the intervention would differ only slightly from those obtained presently.

From 1960 to 1969, the High Aswan dam was constructed with the assistance of the Soviets. The High Aswan dam is much larger than the Aswan Dam that was completed under the supervision of the British in 1902. Lake Nasser, located behind the High Aswan Dam, completely covers the region formerly occupied by the lake formed by the Aswan dam. The effects of the High Aswan dam on the hydrological regime of the Nile River are reported by Shalash (1980a). In an accompanying paper, Shalash (1980b) tabulates the influences of the High Aswan dam on the hydrochemical regime of the Nile River. However, a stochastic tool such as intervention analysis is not employed by Shalash (1980a,b) to rigorously analyze any of the reported findings for the High Aswan dam. Interested readers may wish to obtain the hydrological and hydrochemical data for the Nile River in order to carry out their own intervention analysis studies for the High Aswan dam.

Model Construction

An intervention model for modelling the effects of the construction of the Aswan Dam upon the annual flows of the Nile River, was originally developed by Hipel et al. (1975) while other change-point analyses of the Nile flows have been carried out by MacNeill et al. (1991). However, it is shown here and also by Hipel (1981), how the MAICE procedure from Section 6.3 simplifies the selection of the best model which is more plausible than the model suggested by Hipel et al. (1975). The Nile intervention model can be written using [19.2.8] in the general format as

$$y_t - \mu_y = v(B)S_t^{(T)} + N_t$$

where T stands for October 1, 1902, and the intervention series is represented by

$$\xi_t = S_t^{(T)} = \begin{cases} 0, & t < \text{Oct. } 1, 1902 \\ 1, & t \geq \text{Oct. } 1, 1902 \end{cases}$$

The dynamic and noise components for the intervention are now designed separately.

Designing the Dynamic Component: Based upon a physical understanding of the problem, one would expect the intervention to take place as a step function where the mean drops or steps downwards from 1902 onwards. Figure 19.2.1 confirms that there is a step decrease in the mean level starting at about 1902. This step drop in the mean level is also confirmed by the cusum plot shown for the Nile River in Figure 19.2.9, which is calculated using [19.2.21]. Notice that the cusum graph in Figure 19.2.9 is similar to the one in Figure 19.2.6. The downward sloping ramp from 1902 onwards is caused by the smaller mean level after the intervention.

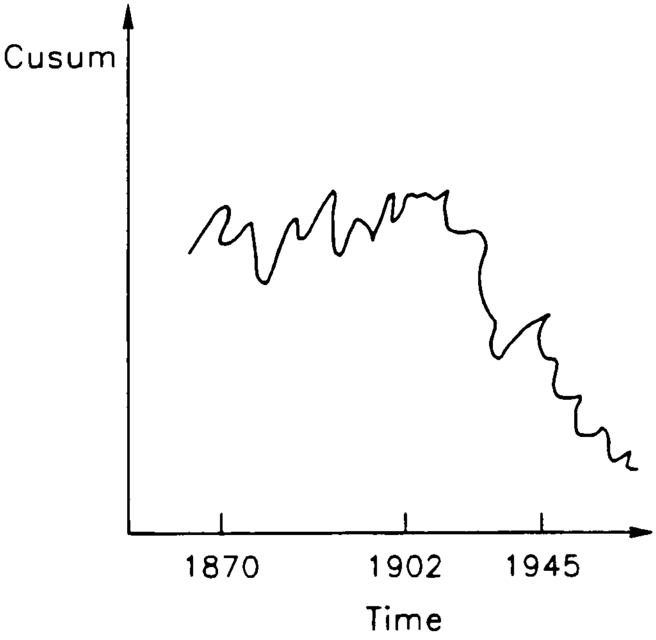


Figure 19.2.9. Cusum for the average annual flows of the Nile River.

By comparing Figures 19.2.1 and 19.2.2b, or, alternatively, relating the properties of Figure 19.2.9 to those of Figure 19.2.6, it can be seen that the component in [19.2.2] for the intervention model can be characterized by a step dynamic response of the form

$$f(\mathbf{k},\xi,t) = \omega_0 S_t^{(T)}$$

One would probably suspect that a transfer function with the parameter ω_0 would be appropriate to reflect the step intervention. However, it is possible that there could be some initial transient effects which require a transfer function of the form $\omega_0/(1 - \delta_1 B)$ where ω_0 and δ_1 are the transfer function parameters. For example, it may take two or three years for the ground water levels to reach a steady-state condition after the reservoir is filled. For this situation, the dynamic component is given as

$$f(\mathbf{k},\xi,t) = \frac{\omega_0}{(1 - \delta_1 B)} S_t^{(T)}$$

where the step function is the same as defined above. By expanding this equation using the binomial theorem as

$$\frac{\omega_0}{1-\delta_1 B} S_t^{(T)} = \omega_0(1 + \delta_1 B + \delta_1^2 B^2 + \cdots) S_t^{(T)}$$

one can appreciate how the transient effects operate. For instance, suppose that t is set equal to 1905. Then the step dynamic response is calculated as

$$f(\mathbf{k},\xi,t) = \omega_0(S_{1905}^{(T)} + \delta_1 S_{1904}^{(T)} + \delta_1^2 S_{1903}^{(T)} + \delta_1^3 S_{1902}^{(T)} + \delta_1^4 S_{1901}^{(T)} + \cdots)$$

$$= \omega_0(1 + \delta_1 + \delta_1^2 + \delta_1^3 + 0)$$

Because $|\delta_1| < 1$ in order for the roots of $1 - \delta_1 B = 0$ to lie outside the unit circle, it can be seen that the transient impacts will disappear after a few years and that the dynamic response will reach the steady state gain from [17.2.3] of $\omega_0/(1 - \delta_1)$. The steady state gain for a step intervention where there is an increasing mean is depicted in Figure 19.2.2d.

Identifying the Noise Component: The noise component is designed by employing the second approach described in Section 19.2.3. Firstly, it is assumed that the noise is white and hence the intervention model has the form

$$y_t - \mu_y = \nu(B) S_t^{(T)} + a_t$$

where $\nu(B)$ can be either ω_0 or $\omega_0/(1 - \delta_1 B)$. Next, the estimates for the innovation sequence, a_t, are obtained along with the MLE's for the model parameters for models with $\nu(B) = \omega_0$ and $\nu(B) = \omega_0/(1 - \delta_1)$. Finally, as expected, the residual series are not white, and are identified following the methods in Chapters 5 to 7, to be either ARMA(1,0) or ARMA(0,1).

MAICE Procedure: Because annual riverflow data sometimes requires a logarithmic transformation, models could be considered where the Box-Cox parameter in [3.4.30] is $\lambda = 0$ for a logarithmic transformation as well as $\lambda = 1$ for no transformation. Of course, other values of λ could also be checked but based on previous modelling experience with riverflow data, only these transformations are considered here. By varying the choice of the Box-Cox parameter λ, $\nu(B)$ and N_t, different models can be considered for modelling the Nile River data. In Table 19.2.1, a range of intervention models are considered for modelling the Nile River time series. For each model, a X entry indicates the type of component contained in the model. Notice that in addition to ARMA(1,0) and ARMA(0,1) noise terms, the white noise ARMA(0,0) model is also included for comparison purposes.

From Table 19.2.1, the minimum value of the AIC occurs for model number 1. The MLE's and standard errors (SE's) given in brackets for this model are listed in Table 19.2.2 while the difference equation for this intervention model is written as

$$y_t - 3340.793 = -715.190\xi_t + (1 + 0.432B)a_t \qquad [19.2.23]$$

From Figure 19.2.10, a plot of the residual ACF, calculated using [7.3.1], reveals that the estimated values fall within the 5 percent significance interval. Hence, the most appropriate intervention model, designed according to the MAICE procedure, possesses residuals that are white. Furthermore, these residuals are approximately normally distributed and homoscedastic.

A comparison of the AIC values in Table 19.2.1 demonstrates that the models which assume an ARMA(0,0) term for N_t (i.e., models 3, 6, 9 and 12) are much less desirable than the

Table 19.2.1. Intervention models for the Nile River.

Model Number	Box-Cox Parameter λ		Transfer Function $v(B)$		Noise Term N_t			AIC
	1.0	0.0	ω_0	$\frac{\omega_0}{1-\delta_1 B}$	ARMA (0,1)	ARMA (1,0)	ARMA (0,0)	
(1)	X		X		X			905.145
(2)	X		X			X		905.800
(3)	X		X				X	941.829
(4)	X			X	X			905.927
(5)	X			X		X		906.001
(6)	X			X			X	943.705
(7)		X	X		X			906.005
(8)		X	X			X		906.366
(9)		X	X				X	941.730
(10)		X		X	X			906.729
(11)		X		X		X		906.468
(12)		X		X			X	943.578

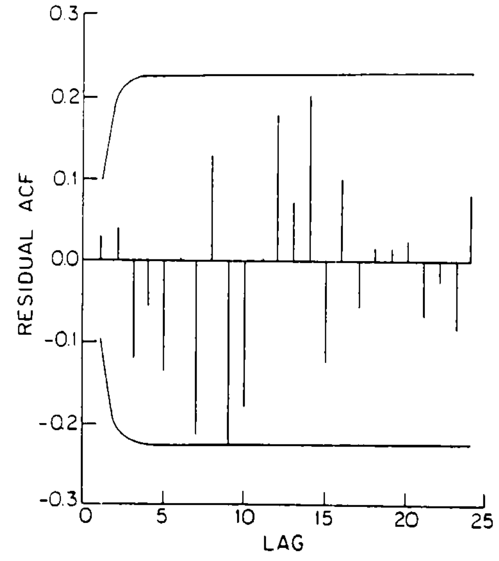


Figure 19.2.10. Residual ACF for the Nile River intervention model.

other models. Whenever an ARMA(0,1) noise term is used instead of an ARMA(1,0) component, it causes an improvement in the AIC value. The AIC entries in Table 19.2.1 also confirm that it is not necessary to take natural logarithms of the data. In addition, a comparison of the AIC values between models 1 and 4 reveals that the type of transfer function causes a difference between the AIC values of less than unity. Although a transfer function of the form ω_0 is more preferred, both from a physical understanding of the problem and also the MAICE

Table 19.2.2. Parameter estimates for the best Nile intervention model.

Parameter	MLE (Standard Error)
ω_0	-715.190 (130.872)
θ_1	-0.432 (0.1041)
μ_y	3340.793 (66.247)
σ_a^2	1.605×10^5

procedure, the fact of the matter is that $\omega_0/(1-\delta_1 B)$ in model 4 is not radically different from ω_0 in model 1. When the parameter estimates are substituted into the aforesaid two transfer functions, the steady-state gains for both models are quite close. Finally, when the MAICE procedure is not invoked, an inferior model may be chosen. Hipel et al. (1975) suggested that model 8 be selected to model the Nile River while the results from Table 19.2.1 can be used in [6.3.2] to show that the plausibility of model 8 versus model 1 is 0.543.

Effects of the Intervention

The model in [19.2.23] can be used for applications such as forecasting and simulation. However, by following the development of [19.2.13] and [19.2.14] in Section 19.2.2, the intervention model can also be employed to statistically describe the change in the mean level of the Nile River due to the Aswan dam. By subtracting the expected value of y_t in [19.2.23] after the intervention from the expected value of y_t before 1902, the drop in the mean level is obtained from [19.2.15] as $-\hat{\omega}_0 = 715.19\, m^3/3$. The percentage change in the mean level is calculated from [19.2.16] to be -21.41% where $\mu_y = 3340.793$ and $\hat{\omega}_0 = -715.19$ from Table 19.2.2 are substituted into the equation. The 95% confidence limits can be determined by adding to and subtracting from $\hat{\omega}_0$, 1.96 times its SE of 130.872. These limits show that the change in the average flows is probably not greater than 971.699 m^3/s and not less than 458.681 m^3/s. By substituting each of these values into [19.2.16] in place of ω_0 and using the estimate of $\mu_y = 3340.793$ for μ_y, the 95% confidence interval for the percentage decrease in the mean flows is from 13.73 to 29.09 percent while the best estimate of the percentage drop in the average is 21.41%.

19.2.5 Stochastic Influence of Reservoir Operation on the Average Monthly Flows of the South Saskatchewan River

Case Study Description

The South Saskatchewan (abbreviated as S. Sask.) River originates in the Rocky Mountains and flows eastward on the Canadian prairies across the province of Alberta to Saskatchewan, where it joins the North Saskatchewan River northwest of the city of Saskatoon. These two rivers form the Saskatchewan River which flows into Lake Winnipeg in Manitoba, which in turn drains via the Nelson River into Hudson Bay. The area of the basin drained by the S. Sask. River at Saskatoon in 139,600 km^2. In January 1969, the Gardiner dam, which impounds Lake Diefenbaker, came into full operation upstream from Saskatoon on the S. Sask. River.

Before the creation of Lake Diefenbaker, the S. Sask. River at Saskatoon usually had higher flows from April to August, with declining flows during the fall and low flows in the winter. The worst floods occurred in the summer when rainfall coincided with heavy snow melt flows from the mountains.

In July 1958, the Canadian and Saskatchewan governments agreed to construct the S. Sask. River project (Saskatchewan Government, 1974). This undertaking consisted of a large dam, spillway and diversion tunnels known as the Gardiner Dam, as well as a much smaller dam and diversion conduit known as the Qu'Appelle Valley Dam. Releases through the latter dam to the Qu'Appelle River represent less than 1% of the flow of the S. Sask. River. Lake Diefenbaker was formed behind these dams. The Coteau Creek generating station was constructed at the Gardiner dam by the Saskatchewan Power Corporation. The East Side pumping station was built at the Gardiner dam to withdraw water for irrigation developments near Outlook and for the Saskatoon-Southeast water supply system.

The downstream flows of the S. Sask. River were not affected by the dam construction until 1964. Part of the water was stored between 1965 and 1969 as the construction neared completion. During the filling period, flows were maintained downstream by releasing water through the diversion tunnels. From September 1968, these releases were used for power generation at the Coteau Creek generating station. Full reservoir operation commenced in 1969. Corrections have been made to the monthly flows at Saskatoon to allow for the effects of various construction phases from 1964 onwards. Because full operation was started in 1969 and the exact construction schedule is not readily available, corrected flows are used from January 1964 to December 1968 in the intervention analysis. These corrected flows represent the flows that would have occurred at Saskatoon if the dam were not being built. The actual flows measured at Saskatoon are used from January 1942 to December 1963 and also from 1969 to 1974 inclusive.

When filled to capacity, Lake Diefenbaker covers an area of 430 km^2 and contains 9.40 km^3 of water. About 308 km^2 are permanently flooded with 5.50 km^3 of permanent storage. This leaves 3.90 km^3 available for flow regulation. The lake is filled each spring and summer when flows are high and water is released during the fall and winter. This type of operation is essential for providing reliable flows throughout the year for power generation at the Coteau Creek generating station.

Besides power generation, the reservoir provides other valuable benefits to the community. The magnitude of floods have been lessened and conversely, minimum flows downstream are guaranteed throughout the year. The inhabitants have taken advantage of the recreational benefits of such a large body of water. Consumptive uses include irrigation and municipal and industrial water supply. Although these consumptive benefits are important, they utilize only a small fraction of the total flow of the S. Sask. River.

Fortunately, most of the uses of Lake Diefenbaker are compatible with the release schedule. During the summer, the reservoir is filled by large flows from the snow melt in the Rocky Mountains. Furthermore, flood extremes are reduced, there is sufficient excess flow for power generation, irrigation and maintenance of minimum downstream flows and the water levels in the lake are high, allowing for optimum recreational benefits. In the winter, the water level is lowered to meet peak power demands and at this time of year recreational requirements are at a minimum. By lowering the reservoir in winter, storage space is available for flood flows which occur in the following year. Consumptive uses require only a small portion of the total flow and therefore are satisfied throughout the year.

The total annual volume of water that flows to Saskatoon is decreased because of losses to consumptive uses through the East Side pumping station and because of releases to the Qu'Appelle Valley. However, the largest loss of water results from natural causes due to the creation of the reservoir. Evaporation losses are high in the summer as a result of the arid climate. Seepage losses are also great but are expected to decrease as groundwater in the area adjusts to the new conditions.

There is no doubt that the Lake Diefenbaker project has significantly altered the flow patterns of the monthly flows of the S. Sask. River at Saskatoon. Downstream users would be interested in the change in mean levels at different times of the year. A decrease in maximum flows is required for flood control and the maintenance of minimum levels is necessary for aquatic life, ferry crossings and adjacent docking facilities, water supply inlets and other appropriate reasons. Therefore, a useful application of intervention analysis is to determine the statistical alteration of the average monthly flows due to the operation of the Gardiner dam. Besides describing the intervention effects, the intervention model can also be used for applications such as simulation and forecasting. The intervention analysis study presented in this section follows the research results of Hipel et al. (1977a).

Model Development

The operation of the Gardiner dam and storage capabilities of the Lake Diefenbaker reservoir changed the previous flow patterns of the S. Sask. River at Saskatoon. As illustrated in Figure 19.2.11, noticeable changes occur subsequent to January 1969. After the reservoir intervention, flows were lowered during the spring and summer and increased during the winter time as compared to before dam construction. Both a cusum chart and monthly plot for each month of the year confirmed these changes (see Section 19.2.3 for a description of how to construct these graphs). These graphs suggested that flows were increased in the months of November to March, inclusive, decreased during April to September and remained about the same in October. It also was evident that the changes occurred as either step increases or decreases.

Designing the Dynamic Component: For seasonal riverflow data, taking natural logarithms of the data is usually a reasonable transformation to invoke for removing heteroscedasticity and non-normality of the residuals. Therefore, based upon an engineering knowledge of the situation and the information from the identification procedures, a possible model for the dynamic component is

$$y_t - \bar{y} = \sum_{i=1}^{12} \omega_{0i} \xi_{ti}$$

where $y_t = \ln Y_t$, natural logarithms of the S. Sask. River monthly riverflows at Saskatoon; $\bar{y}$ is the mean of the entire y_t series;

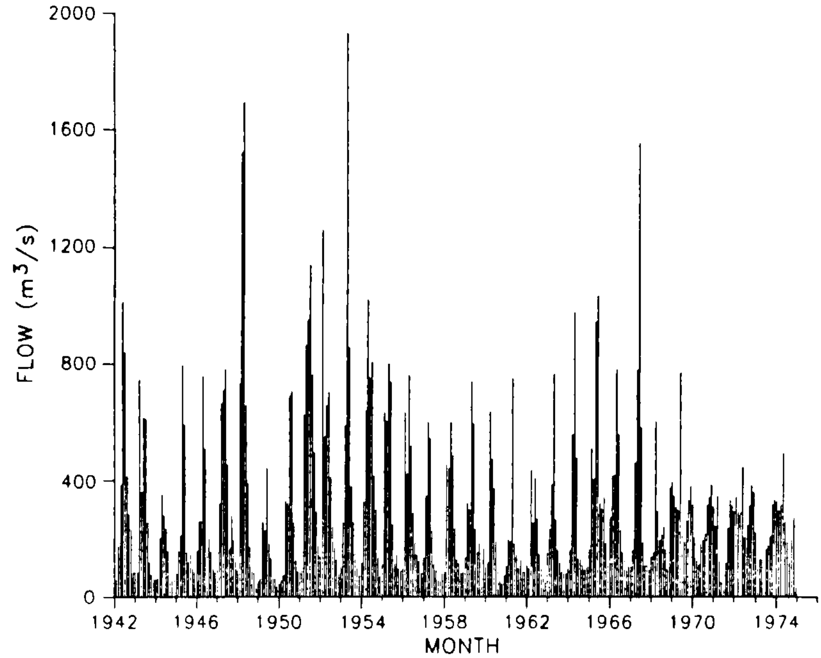


Figure 19.2.11. Average monthly flows of the S. Sask. River.

$$\xi_{ti} = \begin{cases} 1, & t = ith\ month\ \text{for}\ all\ years\ after\ 1968 \\ 0, & otherwise \end{cases}$$

the intervention time series for the ith month of the year where January is considered the first month and December the twelfth month; and, ω_{0i} is the transfer function parameter for the ith month.

Identifying the Noise Component: The noise term is given by

$$y_t - \bar{y} - \left[\sum_{i=1}^{12} \omega_{0i} \xi_{ti} \right] = N_t$$

In order to identify N_t, initially it can be assumed that N_t is white noise. After fitting the resulting intervention model to the logarithmic data from January, 1942, to December, 1974, the form of the SARMA or SARIMA model (see Chapter 12) required for modelling N_t can be identified by examining the residuals using the techniques in Chapters 5 to 7. The ACF of the residuals do not decrease in value for increasing lags that are integer multiples of 12. This indicates that seasonal differencing defined in [12.3.2] may be necessary.

If seasonal differencing is used, this indicates that the series is nonstationary and does not fluctuate about any mean level. However, as discussed in Part VI and elsewhere, it is known that for seasonal hydrological time series, for which the effects of any interventions are suitably accounted for, the observations within each season tend to fluctuate about an overall mean level and are, therefore, seasonally stationary. Consequently, for the application of intervention analysis considered here for average monthly riverflows, differencing is not desirable. In order to rectify the situation, a deterministic component is brought into the model. The average monthly logarithmic flows for each month of the year before 1964 are calculated. Recall that January 1964 was the time that dam construction started and corrected flows are used from 1964 to 1968. The monthly logarithmic average for each month is subtracted from the natural logarithm of that month for each year from 1942 to 1974. In other words, the logarithmic data are deseasonalized using [13.2.2].

Following deseasonalization, the deseasonalized flows are used in the above intervention model where it is first assumed that N_t is white. An ARMA model to fit to the residuals is then identified. The graphs of the residual ACF, PACF, IACF and IPACF and their 95% confidence limits are given in Figures 19.2.12 to 19.2.15, respectively (see Section 5.3 for a discussion of how to construct these graphs). Notice that the PACF and IACF truncate after lag one, while the ACF and IPACF have a large value at the first lag with decreasing magnitudes at larger lags. These facts indicate that an ARMA(1,0) or Markov model can model the noise term as

$$(1-\phi_1 B)N_t = a_t$$

or

$$N_t = \frac{a_t}{1-\phi_1 B}$$

Estimation and Diagnostic Checking: From the identification stage, the model to estimate is:

$$y_t - \bar{y} = x_t + \sum_{i=1}^{12} \omega_{0i}\xi_{ti} + \frac{a_t}{l-\phi_1 B} \qquad [19.2.24]$$

where x_t is the deterministic component formed by 33 consecutive sequences of the twelve monthly means of the natural logarithms of the monthly flows before 1964. Keep in mind that the deterministic component simply means that the logarithmic data are deseasonalized using [13.2.2].

Table 19.2.3 lists the MLE's and SE's for the model parameters in [19.2.24]. Diagnostic checks reveal that the assumptions that the a_t's are independent, homoscedastic and normally distributed, are satisfied. Therefore, based on the data used, the intervention model in [19.2.24] adequately models the operation of the Gardiner dam.

Effects of the Intervention

Because natural logarithms were taken of the response variable in [19.2.24], in order to express the transfer function parameter in terms of the original data, a transformation must be calculated. The following calculations are similar to those executed in Section 19.2.2 under the heading "Example with a Logarithmic Data Transformation and a Step Intervention". Taking natural antilogarithms of [19.2.24] gives

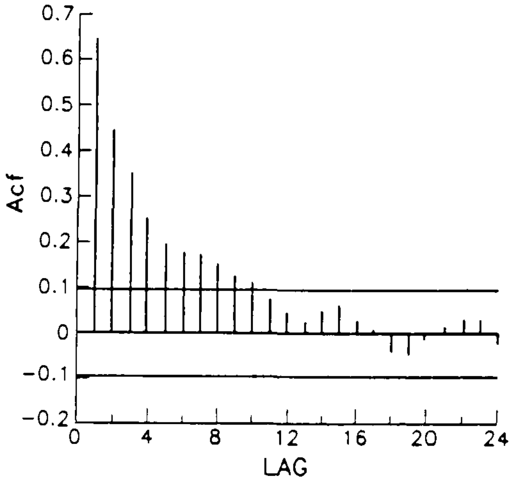


Figure 19.2.12. ACF and 95% confidence limits for the S. Sask. River residuals.

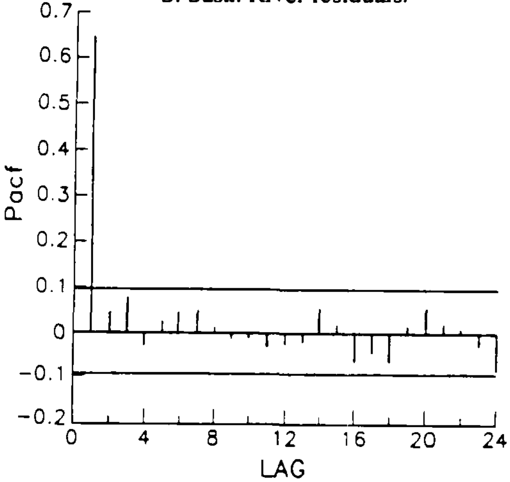


Figure 19.2.13. PACF and 95% confidence limits for the S. Sask. River residuals.

$$\begin{aligned} y_t &= e^{\bar{y}} e^{x_t} e^{N_t} \exp\left[\sum_{i=1}^{12} \omega_{0i} \xi_{it}\right] \\ &= c_1 e^{x_t} e^{N_t} \exp\left[\sum_{i=1}^{12} \omega_{0i} \xi_{it}\right] \end{aligned}$$

where $c_1 = e^{\bar{y}}$ is a constant.

Before the dam came into full operation in January 1969, the intervention time series have values of zero. Thus, taking expectations, the above equation gives

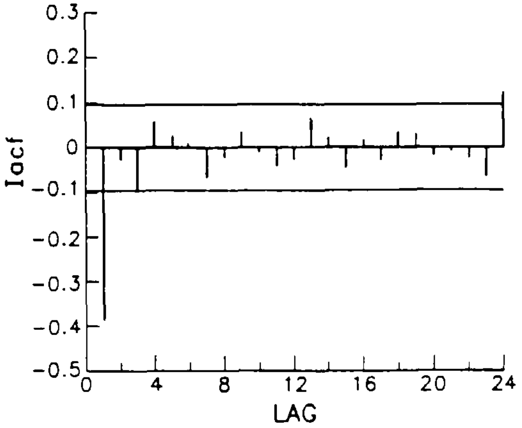


Figure 19.2.14. IACF and 95% confidence limits for the S. Sask. River residuals.

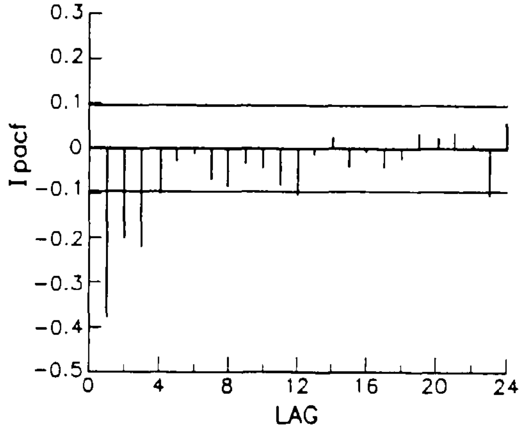


Figure 19.2.15. IPACF and 95% confidence for the S. Sask. River residuals.

$$E[Y_t]_{before} = c_1 c_2$$

where

$$c_2 = E\left[e^{x_t}e^{N_t}\right]$$

For each year after 1968, ξ_{ti} is unity for the ith month and zero otherwise. The expected value of Y_t for month i after 1968 is:

Table 19.2.3. MLE's for the parameters in the S. Sask. intervention model.

Parameter	Estimate	Standard Error
ω_{01} (Jan.)	1.673	0.172
ω_{02} (Feb.)	1.602	0.181
ω_{03} (Mar.)	1.041	0.185
ω_{04} (Apr.)	-0.541	0.186
ω_{05} (May)	-0.698	0.187
ω_{06} (June)	-1.002	0.188
ω_{07} (July)	-0.731	0.188
ω_{08} (Aug.)	-0.431	0.188
ω_{09} (Sep.)	-0.212	0.187
ω_{010} (Oct.)	0.176	0.186
ω_{011} (Nov.)	0.744	0.184
ω_{012} (Dec.)	1.441	0.179
ϕ_1	0.651	0.038

$$E[Y_t]_{after} = c_1 c_2 e^{\omega_{0i}}$$

Utilizing the foregoing, the percentage change in the mean level of the flow for month i due to the intervention is:

$$\%\ change = \left(\frac{E[Y_t]_{after}}{E[Y_t]_{before}} - 1 \right) 100 = (e^{\omega_{0i}} - 1)100 \qquad [19.2.25]$$

Interpretation of Results

The operation of the Gardiner dam significantly affected the average monthly flows of the S. Sask. River at Saskatoon. An examination of the transfer function parameter estimates in the second column of Table 19.2.3 and the corresponding SE's in the third column indicates which changes are significant. As was suspected, there are significant increases in the flows from November to March. Conversely, as indicated by the negative signs, the average flows decrease from April to September. Because the MLE's possess a limiting normal distribution, hypothesis testing can be done. Notice that the transfer function parameter estimate for September is not significantly different from zero for a one sided test at the 10% significance level. The October parameter estimate shows a slight increase but this is not significantly different from zero since the SE is greater than ω_{010}.

By substituting the transfer function parameter estimate for each month into [19.2.25], the estimate can be transformed into percentage change in flow. Table 19.2.4 lists the average monthly flows before 1964 and the percentage change in mean monthly flow from 1969 to 1974. For any month i, confidence limits can be calculated for the percentage alteration in mean level.

Table 19.2.4. Average monthly flows for the S. Sask. River before reservoir operation and the percentage changes from 1969 to 1974.

Month	Average Flow Before 1964 (m^3/s)	Percentage Change
Jan.	68.69	432.86
Feb.	69.71	396.31
Mar.	71.91	183.30
Apr.	393.98	-41.80
May	425.72	-50.25
June	790.51	-63.29
July	595.56	-51.89
Aug.	285.98	-35.00
Sep.	228.24	-19.10
Oct.	169.22	16.17
Nov.	120.10	110.52
Dec.	79.42	322.55

The 95% confidence limits are determined by adding to ω_{0i} and subtracting from ω_{0i} 1.96 times its SE and substituting these two values into [19.2.25] in place of ω_{0i}. For example, the 95% confidence limits for January indicate that very likely the increase in average is not greater than 646% and not less than 281%. The best estimate of this increase is 433%. This type of statistical description of the mean flow changes is only possible by using the intervention analysis technique.

If the new mean level for January is required in m^3/s, simply multiply 68.69 times (4.3286 + 1) to obtain 366.02. The 95% confidence interval for the January mean flow after the intervention is (261.43, 512.46). It should be noted that the arithmetic average for six January flows after 1968 is 354. This is very close to the value of 366 obtained by intervention analysis and is within the 95% confidence interval of the January average flow after reservoir operation started.

Intervention analysis is a viable technique to model and statistically describe the effects of reservoir operation on the downstream flows from a dam. For the particular problem analyzed in this section, the percentage changes of the average monthly flows of the S. Sask. River at Saskatoon due to reservoir operation, are determined. Although the mean flow changes are calculated separately for each month, for other applications it is possible to analyze changes for specific sets of months. For instance, a certain problem may deem it necessary to calculate the changes in flow patterns over a whole season, such as for the summer, or winter months, rather than for each individual month. Intervention analysis may also be used to test whether or not a change in operating rules of a dam already in operation, significantly alters average flows. Of course, in addition to descriptive purposes, any intervention model developed can also be used for forecasting and simulation.

Notice in [19.2.25], that a separate intervention component is developed for each month. One may wonder if a separate noise model should also be estimated for each month or season. In other words, in a fashion similar to a periodic seasonal model in Chapter 14, a periodic intervention model could be developed where there is in effect a separate intervention model for each season. This is precisely what is done in Section 19.6 for the S. Sask. River data. As is shown, the results obtained are close to those in Tables 19.2.3 and 19.2.4 for the model in [19.2.24].

19.3 DATA FILLING USING INTERVENTION ANALYSIS

19.3.1 Introduction

An assumption underlying virtually all of the time series models which can be employed in practical applications is that the data sets to which they are fitted consist of observations separated by equal time intervals. Although it would be desirable to possess stochastic models which can readily handle time series consisting of any kind of unevenly spaced observations, currently no such practical models exist and, indeed, it may turn out to be mathematically intractable to develop these types of stochastic models. In practice, if the measurements are not evenly spaced, appropriate techniques must be utilized to produce a series of equally spaced data that is estimated from the given information. Of course, as explained in Section 19.7 and also by Lettenmaier et al. (1978), practitioners are advised to design future sampling programs so that evenly spaced data are collected at suitable time intervals. In this way, the inherent assets of available time series models, such as those discussed throughout this book, can be fully exploited.

Time series with missing observations or, equivalently, time series where the measurements are taken at unequal time intervals, occur quite often in practice in various fields. For instance, as noted by authors such as Hirsch et al. (1982), McLeod et al. (1983) and D'Astous and Hipel (1979), as pointed out in Section 1.2.4 and throughout Part X, and as demonstrated by the applications in Sections 19.3.6, 19.4.5, 22.4.2, 23.5.2 and 24.3.2, the problem of missing values in data sequences happens frequently in environmental engineering. There are many reasons why environmental data are often not collected at evenly spaced points in time. Sometimes bad weather conditions make it difficult to collect the data. As noted by D'Astous and Hipel (1979), water quality data cannot be collected sometimes during the winter time when the ice on lakes and rivers is too thick. Likewise, Baracos et al. (1981) mention that hydrometeorological records from the Arctic regions often contain missing observations due to the breakdown of equipment which cannot be repaired when severe climatic conditions make the measuring station inaccessible.

Another reason for not obtaining evenly spaced measurements is that there are conflicting demands regarding how the data will be used and hence how it should be collected. Because all the fish in a lake will die if the dissolved oxygen level goes to zero only once, a biologist may wish to take many dissolved oxygen measurements whenever the critical value of zero is approached whereas when it is suspected that there is sufficient dissolved oxygen he may not require very many observations. On the other hand, a scientist who wishes to use intervention analysis for modelling trends caused by external interventions requires that an equally spaced time series be available. Of course, if a properly designed sampling procedure is implemented both demands can be satisfied by taking frequent measurements during the critical periods when the dissolved oxygen is low and by taking equally spaced observations at other times. From this data base, an equally spaced series can be conveniently and efficiently estimated.

In addition to conflicting demands, there is another reason why environmental as well as other types of data are often not properly collected. In many countries, certain agencies are responsible for collecting the data and other institutions are committed to analyzing the time series. Because the collection agencies may not be aware of the analytical tools that will eventually be employed for detecting valuable information in the data, they often adopt incorrect sampling procedures. Only when the mathematical characteristics of the analytical tools are taken into account, can an appropriate data collection scheme be devised (Lettenmaier et al., 1978). Whatever the reasons, time series often contain unequally spaced data and techniques are required for efficiently estimating the missing observations.

The main purpose of this section is to present an efficient data filling technique which is actually a special kind of intervention model. In Section 19.4 it is explained how multiple interventions and estimating missing observations can be simultaneously handled using an intervention model, while in Section 19.5 multiple input series are also included in order to form the most general case of the intervention model. However, within Section 19.3 it is assumed that there are no external interventions and an intervention model is designed for estimating missing observations where up to about 10% of the data may not be recorded. Prior to defining the special type of intervention model and demonstrating how it is used for data filling, existing techniques for creating an equally spaced time series are discussed next.

19.3.2 Techniques for Data Filling

Data Filling Methods Presented in this Text

Within this book, three different procedures are given for estimating missing observations. The techniques are specifically designed for data filling in different types of situations which can arise in practice and are briefly outlined below.

1. **Back Forecasting:** The first approach which is discussed in detail in Section 18.5.2 is referred to as back forecasting and can be used to extend hydrometeorological records. For example, as noted by Baracos et al. (1981), meteorological measurements such as temperature and precipitation have been kept in the Canadian Arctic for a much longer time than riverflow series. For the data where the riverflow and meteorological series intersect in time, a TFN model can be built following the procedures of Chapter 17 to obtain a model with the riverflows as the single output and the covariates such as precipitation and temperature as the inputs. Using this TFN model and the meteorological data which do not overlap in time with the riverflows, the earlier unknown measurements for the riverflows can be back forecasted. Beauchamp et al. (1989) follow a similar procedure for extending daily flows in a river based upon a TFN model that connects these flows to longer upstream records. Finally, Grygier et al. (1989) present another approach for extending correlated series.

2. **Intervention Analysis:** The second technique employs a special form of the intervention model to efficiently estimate missing data points when not more than about 10% of the data are missing. This procedure is described in detail in Section 19.3.3 and also used with the other kinds of intervention models discussed in Sections 19.4 and 19.5. Besides the applications given in Sections 19.3.5, 19.3.6, 19.4.4 and 19.4.5 of Chapter 19, intervention analysis is utilized for estimating missing values in examples presented within Section 22.4.2 of Chapter 22. In essence, the intervention analysis approach to data filling is equivalent to the method presented by Coons (1957) which was originally given in a paper by Bartlett (1937) and also described by Anderson (1946). The data filling method described by Coons (1957) can be used when one or more missing observations exist in an

experiment of any statistical design where the errors are assumed to be normally and independently distributed. As noted by Coons (1957), the advantages of this method are its generality of application and the ease with which exact tests of significance may be obtained. When this general approach is utilized within the framework of the intervention model, a flexible data filling technique can be constructed.

3. **Seasonal Adjustment:** When dealing with some types of time series, especially environmental data, often there are many missing data points where there may be long periods of time for which no observations were taken. In addition, there may be one or more external interventions which cause trends in the series. To estimate the many missing observations for this *messy* type of data, a procedure based on seasonal adjustment can be employed. In Section 22.2 the seasonal adjustment technique is formulated and used to reconstruct water quality time series in the applications in Chapter 22.

Additional Data Filling Methods

A variety of approaches to *data interpolation* is described in the published literature. For example, Wilkinson (1958) and Preece (1971) deal with estimating missing values for experimental data. Specially designed regression models can be designed for estimating missing values in a data sequence. For example, the robust locally weighted regression smooth devised by Cleveland (1979) and described in Section 24.2.2, could be employed for data filling. Using both a regression analysis model and TFN model that connects upstream and downstream daily flows in a river, Beauchamp et al. (1989) extend the shorter downstream records. Nonetheless, as pointed out at the end of Section 17.2.4 and also by Beauchamp et al. (1989), regression models possess a structure which is not as general as the TFN models of Chapter 17 or intervention models of this chapter, since the noise terms in regression models are assumed to be white rather than correlated and the transfer functions are not as well formulated. Consequently, Beauchamp et al. (1989) recommend using a TFN model for record extensions. Regression and other kinds of formal models can be used in conjunction with graphical displays of the series being studied to fill in missing values. However, data filling methods which do not explicitly take the autocorrelation structure of a series into account, are not properly designed for use with time series data.

Brubacher and Wilson (1976) have devised a technique that is an application of the least squares principle and forecasting approach to estimate the effect of one-day national holidays on hourly electricity demand. This is done by interpolating over the holiday period using unaffected electricity demand observations from both before and after this period. The interpolated values are obtained through a method that makes use of forecasting and back-forecasting procedures to regenerate the residual series. The interpolates are then determined so as to minimize the sum of squares of these regenerated residuals. This estimation technique leads to a set of k equations to be solved for k interpolates. The ratio of the actual demand to the estimated or interpolated normal demand, recorded for the same holiday period over successive years, may then be employed to forecast the effect on future holiday demands.

The interpolation technique developed by Brubacher and Wilson (1976) seems to be adequate for the application in question but is fairly complex even if very few missing values must be estimated. The nature of the electricity demand data is such that an appropriate ARIMA model representing the whole time series can be identified from a subset of the data. This is because the yearly patterns of the series are insignificant so that modelling the weekly patterns is adequate. For instance, only four or five weeks of data provide sufficient information to identify a suitable model. Consequently, the effect of the holiday does not create a problem in finding an adequate model. There are enough data before and after the given holiday period to justify the use of the selected ARIMA model for forecasting and back forecasting the interpolates. However, in practice, the interpolation technique of Brubacher and Wilson (1976) is not so readily applicable to most time series. If many observations are missing, it becomes increasingly difficult to select a proper model for the time series. The reliability of the forecasted values is also a function of the number of gaps in the data. Another factor to consider is the proximity of the data gaps to the beginning or end of the time series. For example, if a missing data point were in the middle of the series, there may be insufficient data either before or after the gap to formulate an adequate forecasting model. The forecasted or back forecasted interpolate is therefore not dependable. Furthermore, if the data to be interpolated are subjected to one or more external interventions, then most ARIMA models are not suitable and forecasts should not be based with these models. These arguments imply that this method of data filling is not admissible for data that has been affected by known external interventions.

Other research related to the problem of missing observations can also be found in the literature. For instance, Marshall (1980) devises a technique for estimating the ACF of a time series when there are missing observations which are assumed to occur randomly. Within the frequency domain, a number of authors have considered problems which arise in spectral analysis when observations are missing at random (Jones, 1962; Parzen, 1963; Scheinok, 1965; Bloomfield, 1970; Neave, 1970). The intervention analysis technique to data filling does not assume that the missing data points occur randomly. Finally, Chin (1988) presents a spectral analysis approach to fill in data at one location based on measured data at an adjacent location.

A general approach to iterative computation of MLE's when the observations can be viewed as incomplete data is given by Dempster et al. (1977). Because each iteration of the algorithm consists of an expectation step followed by a maximization step, the authors call it the *EM algorithm*. This procedure is ideal for estimating simultaneously both missing values and the parameters of the model being fitted to the data set. As a matter of fact, the EM algorithm could be employed in conjunction with the intervention models for data filling defined in Sections 19.3.3, 19.4.2 and 19.5.2. At each iteration, the missing values are replaced by their expectation given the current parameter values (called the E-step) and then the parameters are estimated once again (M-step). The iterations are continued until the estimates exhibit no important changes.

Based upon a state space formulation, Jones (1980) develops a maximum likelihood estimator for fitting ARMA models to time series having missing observations. Additionally, Ljung (1982) develops an expression for the likelihood function of an ARMA model when some observations are missing and shows how the missing data points can be estimated from the available data. Finally, Little and Rubin (1987) describe a wide range of approaches for dealing with missing data.

19.3.3 Model Description

Suppose that there are no external interventions which are affecting a given series which has missing observations. When the number of missing data points is not excessive, the intervention model can be employed for data filling. Qualitatively, an intervention model for handling this situation can be written as *response variable = dynamic component + noise*

where the dynamic component contains intervention terms which can be used for estimating the missing data points. In a more precise fashion, an intervention model for modelling a series with multiple missing data points can be described by

$$(y_t - \mu_y) = f(\mathbf{k},\xi,t) + N_t \qquad [19.3.1]$$

where t represents discrete time, y_t is the response series which may be transformed using the Box-Cox power transformation in [3.4.30], μ_y is the theoretical mean of the y_t series, N_t is the noise term which is usually correlated and can be modelled using an ARMA or ARIMA model, and $f(\mathbf{k},\xi,t)$ is the dynamic component with a set of parameters, $\mathbf{k}$, and a set of intervention series, ξ. As will be explained, whenever a term in the dynamic component is used to model a missing observation, a specific type of transfer function and intervention series is always used. However, the design of the noise term, N_t, is not fixed and the parameters required in the ARMA representation of N_t must be decided upon in each application. An ARMA model for the noise component is given in [19.2.7].

To specify exactly the form of the model in [19.3.1] where there are no external interventions, first consider the case where there is one missing observation at time t_1, and the response series is not transformed using a Box-Cox transformation defined in [3.4.30]. The intervention model for estimating the missing observation is written as

$$y_t - \mu_y = \omega_{01}\xi_{t1} + N_t \qquad [19.3.2]$$

where ω_{01} is the only parameter in the transfer function, and ξ_{t1} is the pulse intervention series which is set to unity at time $t = t_1$ and given a value of zero elsewhere. Although the missing observation at time t_1 can be assigned any fixed value, it is convenient to assign y_{t_1} a value of zero. After setting y_{t_1} to zero, at time $t = t_1$, the intervention model from [19.3.2] is given as

$$-\omega_{01} = \mu_y + N_{t_1} \qquad [19.3.3]$$

where μ_y can be efficiently estimated by the series mean $\bar{y}$. Notice that the right hand side of [19.3.3] consists of the mean level of the series plus the autocorrelated noise. This in fact is the value of the series at $t = t_1$. Consequently, the MLE for $-\omega_{01}$ constitutes an efficient estimate for the missing value of y_{t_1} where the autocorrelation structure of the series is automatically taken into account in [19.3.3].

Suppose that the y_t series in [19.3.3] requires a Box-Cox transformation to eliminate non-normality and/or heteroscedasticity in the model residuals contained in the noise component, N_t. Then a non-negative value other than zero would have to be initially used for the missing y_t observation at time t_1. Suppose that this value is represented as $\bar{y}_{t_1}$ where, for instance, $\bar{y}_{t_1}$ may simply be the mean of the known transformed observations. At time t_1, the estimate for the missing observation in the transformed domain would be

$$-\omega_{01} = \mu_y + N_{t_1} - \bar{y}_t \qquad [19.3.4]$$

To determine the estimate of the missing observation in the untransformed domain, one would simply take the inverse Box-Cox transformation of $-\omega_{01}$ in [19.3.4].

The model may be expanded to handle a situation where there is more than one missing observation. If I_2 values are missing and there are no external interventions, the model is given as

$$y_t - \bar{y} = \sum_{j=1}^{I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.3.5]$$

where ω_{0j} is the parameter of the jth transfer function and ξ_{tj} is the jth intervention series which is assigned a value of unity where the jth observation is missing and zero elsewhere. If the missing observation at time t_j is initially considered to be zero, then at $t = t_j$, equation [19.3.5] becomes

$$-\omega_{0j} = \bar{y} + N_{t_j} \qquad [19.3.6]$$

Therefore, an efficient estimate for y_{t_j} is the MLE of $-\omega_{0j}$. If the series were transformed using a Box-Cox transformation, then the inverse Box-Cox transformation of the estimate for each missing data point must be taken to obtain the estimate for each missing observation in the untransformed space.

The intervention analysis approach to data filling possesses many inherent attributes. Firstly, as noted earlier, an efficient estimate is obtained for each missing observation along with its standard error of estimation. Because the MLE for each missing data point is known to be asymptotically normally distributed, confidence limits can be calculated for each estimated missing value. Secondly, a moderate number of missing data points can be simultaneously estimated along with the other model parameters. It should be pointed out that the missing data can be estimated at any location in the series, including the initial and final points. Thirdly, as explained in Sections 19.4 and 19.5, intervention analysis can be used to estimate missing observations even when there are multiple external interventions and multiple input series. Finally, as shown in the next section, an intervention model for filling in data can be conveniently constructed by adhering to the identification, estimation and diagnostic check stages of model development. Authors who have employed the intervention analysis approach to data filling within water resources and environmental engineering include D'Astous and Hipel (1979), Lettenmaier (1980) and Hipel and McLeod (1989).

19.3.4 Model Construction

When there are no external interventions and only missing data points, the form of each intervention term in the dynamic component is fixed. For instance, the intervention term for the jth missing observation is

$$v_j(B)\xi_{tj} = \omega_{0j}\xi_{tj}$$

where ω_{0j} is the only transfer function parameter and ξ_{tj} is the jth intervention series which is given a value of one where the jth observation is missing and zero elsewhere. Accordingly, it is only necessary to ascertain the parameters required in the ARMA formulation of N_t.

To design the form of N_t, one of the following techniques can be used where the third method is probably the simplest to use in most situations.

1. First replace each missing value by a "rough" estimate of what it may be. Next, using the entire reconstructed series, identify the form of the ARMA model needed to describe it by following the usual procedures in Chapters 5 to 7. Rough estimates can be obtained in a number of ways where only a simple procedure should be chosen. For instance, each missing observation can be replaced by the mean of the known observations. When the data are seasonal, always replace the missing value by its seasonal mean. Another simple technique is to plot the entire series and visually interpolate among the plotted observations to obtain a rough estimate for each missing observation.
2. If there is a sufficiently long section of data for which there are no missing observations, use this interval of data to identify the form of N_t. Once again, the standard techniques of Chapters 5 to 7 can be used.
3. The third technique is the empirical identification technique presented in Sections 19.2.3, 17.3.1 and 17.5.3. After fixing the form of each intervention term in the dynamic component, fit the model in [19.3.5] to the series where it is assumed that the noise term is white, and, therefore the intervention model in [19.3.5] has the form

   $$a_t = (y_t - \bar{y}) - \sum_{j=1}^{I_2} \omega_{0j}\xi_{tj}$$

   In practice, usually the noise term is correlated. Consequently, after obtaining the estimated residual series, $\hat{a}_t$, for the above model, the kind of ARMA model to fit to the noise series can be determined by following the model development stages given in Chapters 5 to 7.

By using the identified form of N_t along with the fixed format of the dynamic component, an overall design for the model in [19.3.5] is now available. Before estimating the model parameters, each missing data point is initially assigned a value of zero or some appropriate position value. Of course, if the series is first transformed using a Box-Cox transformation, the missing values are given their zero values after obtaining the transformed sequence for the known observations. Otherwise they can be assigned a positive value such as the mean of the known observations before taking the Box-Cox transformation. Using the method of maximum likelihood discussed in Appendix A17.1, efficient estimates can be simultaneously obtained for all the model parameters where the estimate for the jth missing observation is $-\hat{\omega}_{0j}$. The adequacy of the fitted model can be checked by utilizing the tests described in Chapter 7, and Sections 17.3.3, 17.5.3 and 19.2.3. Note that if there are problems with the model residuals, only the form of N_t must be redesigned since the format of the dynamic component is fixed.

19.3.5 Experiments to Check the Performance of the Data Filling Method

From a theoretical viewpoint, the intervention model is known to produce efficient estimates for the missing observations (Coons, 1957; Bartlett, 1937). To demonstrate how well the data filling technique works in practice, it is assessed by estimating observations where the actual historical values are known. Consider the average annual flows from 1860 to 1957 for the St. Lawrence River at Ogdensburg, New York. As explained in Sections 3.2.2 and 5.4.2, the most appropriate model to fit to this sequence is a constrained AR(3) model where the second AR parameter is constrained to zero in the equation

$$(1-\phi_1B-\phi_3B^3)(Y_t-\mu_y)=a_t$$

where ϕ_i is the ith AR parameter (see Section 3.2 for a description of AR models), capital Y is used to emphasize that there is no Box-Cox transformation, and μ_y is the mean of the Y_t series. The equation for this model which contains the values of the estimated parameters is given in [3.2.19] and [6.4.2]. Because the model residuals are approximately normally distributed and homoscedastic, it is not necessary to transform the data using a Box-Cox transformation (this is the case where the Box-Cox parameter λ is set equal to one in [3.4.30]).

The St. Lawrence River time series consists of 97 observations and therefore the time t can be considered to go from $t = 1$ to $t = 97$. The proposed data interpolation method is tested by deleting observations at the beginning, the end, and in other locations of the time series. Table 19.3.1 displays the data filling studies for the St. Lawrence River. The time series entries are given in cubic meters per second while $\lambda = 0$ means that natural logarithms are taken of the original data. To illustrate the mathematical structure of the intervention models in Table 19.3.1, the model for test case 4 is written for the times $t = 33$ and $t = 34$, respectively, as

$$-\hat{\omega}_{01}=\bar{Y}+\frac{1}{1-\hat{\phi}_1B-\hat{\phi}_3B^3}\hat{a}_{33}$$

and

$$-\hat{\omega}_{02}=\bar{Y}+\frac{1}{1-\hat{\phi}_1B-\hat{\phi}_3B^3}\hat{a}_{34}$$

in which $\hat{\phi}_i$ is the ith estimated AR parameter, $\bar{Y}$ is the series mean, and $\hat{a}_t$ is the white noise residual at time t.

From Table 19.3.1, the estimated value for the observation is within two SE's of the actual data point for case 1 while all other estimates are within one SE of the true values. In fact, the estimates are quite close to the actual values even in the case where the very first data point is missing. This indicates that the noise term in the intervention model more than adequately accounts for the particular autocorrelation structure of the time series. Although no Box-Cox transformation is required in the original model, natural logarithms (i.e., $\lambda = 0$) of the data are taken for test case 5 in Table 19.3.1. Thus,

$$-\hat{\omega}_{01}=\bar{y}+\frac{1}{1-\hat{\phi}_1B-\hat{\phi}_3B^3}\hat{a}_{25}$$

in which $y_t = \ln(Y_t + 1)$. The constant must be added since the observation Y_t at time t has been set equal to zero. As shown in Table 19.3.1, the estimate $-\hat{\omega}_{01}$ of $-\omega_{01}$, has a value of 8.95. The estimate for the missing observation in the original series is

Table 19.3.1. Estimates for known observations for St. Lawrence River data.

Test Case	Lag of Missing Observation	λ	$-\hat{\omega}_{0j}$	Standard Error	Actual Value, in Cubic Meters per Second
1	94	1	7,724.27	343.25	7,194.00
2	9	1	7,165.11	342.58	7,051.00
	94		7,226.15	342.58	7,194.00
3	1	1	7,708.63	408.22	7,788.00
4	33	1	6,489.97	378.23	6,583.00
	34		6,427.43	378.23	6,583.00
5	25	0	8.95	0.05	7,660.00

$$\hat{Y}_{25} = e^{-\hat{\omega}_{01}} - 1 = 7,703.81$$

This calculated value is close to the historical magnitude of 7,660.00, which is listed in Table 19.3.1.

19.3.6 Estimating Missing Observations in the Average Monthly Lucknow Temperature Data and Middle Fork Riverflows

In Section 17.5.4, TFN noise models are developed where the output is always the average monthly flows of the Saugeen River at Walkerton, Ontario, Canada, and the covariate series consist of precipitation and temperature data sets from two different locations. As shown in Table 17.5.2, for the Lucknow temperature series there are ten missing observations. These gaps in the time series must be filled in before the covariate temperature series can be used in a TFN model. To accomplish this, the intervention model in [19.3.5] can be utilized.

Before fitting the model, the temperature series is first deseasonalized by employing the technique in [13.2.3] where the series is not initially transformed using a Box-Cox transformation. Next, the first identification technique described in Section 19.3.4 is used to determine which parameters are needed in the ARMA noise term. Because the series is deseasonalized, each missing observation is assigned the monthly deseasonalized mean of zero. Then the form of the ARMA model required for modelling the series and hence, N_t, is determined by following the stages of model construction outlined in Chapters 5 to 7. The noise term is identified to be an ARMA(0,4) model with the second and third MA parameters constrained to zero. Consequently, the particular form of the intervention model in [19.3.5] which can be utilized for modelling the deseasonalized Lucknow temperature series is

$$y_t - \mu_y = \sum_{j=1}^{10} \omega_{0j}\xi_{tj} + (1 - \theta_1 B - \theta_4 B^4)a_t \qquad [19.3.7]$$

where ω_{0j} is the parameter in the jth transfer function, ξ_{tj} is the jth pulse intervention series that is assigned a value of unity where the observation is missing and zero elsewhere, and θ_i is the ith MA parameter (see Section 3.3.2 for a definition of a MA(Q) model). After simultaneously estimating all the model parameters in [19.3.7], the adequacy of the fitted model is confirmed by subjecting the residuals to diagnostic checks.

The estimate for the jth missing data point in the deseasonalized series is given by $-\hat{\omega}_{0j}$. To obtain the estimate and standard error for each missing observation in the original series, they must undergo a reverse deseasonalization transformation as in [13.2.3]. In Table 17.5.2, the estimates of the ten missing data points (and their SE's in brackets) and the actual monthly means are presented for the original untransformed series in the second and third column, respectively. Notice that the difference between each estimate and its monthly mean is always less than its SE.

Another application using the intervention model of [19.3.5] to estimate missing values in a monthly riverflow time series is presented in the subsection called the Middle Fork Intervention Model within Section 22.4.2 of Chapter 22. To model the seasonality contained in the natural logarithms of the average monthly flows of the Middle Fork River, a seasonal differencing operator of order one is included in the SARIMA noise term of the intervention model. Consequently, it is not necessary to deseasonalize the logarithmic Middle Fork Riverflows, as is done for the data in this section.

19.4 INTERVENTION MODELS WITH MULTIPLE INTERVENTIONS AND MISSING OBSERVATIONS

19.4.1 Introduction

In Section 19.2, an intervention model is designed for modelling a time series which may be influenced by multiple external interventions while in Section 19.3 a specialized kind of intervention model is described for obtaining efficient estimates of missing values in a data sequence. The purpose of this section is to present an intervention model which can simultaneously handle both the modelling of the effects of multiple external interventions upon the levels of a series and the estimation of missing observations. As noted in the introduction in Section 19.1, a practical example of this problem is given by the graph displayed in Figures 19.1.1 and 1.1.1 of the average monthly phosphorous (in milligrams per litre) for the Speed River, Ontario, Canada. The external intervention which caused the drop in the level starting in February, 1974, (i.e., the 26th data point) was the implementation of conventional phosphorous treatment at the upstream Guelph sewage treatment plant. Besides the drop in the level caused by phosphorous treatment, the blackened circles indicate that there are missing observations both before and after the intervention. As is shown in Section 19.4.5, an intervention model can be conveniently constructed for modelling the effects of the intervention and obtaining efficient estimates of the missing data for the phosphorous series in Figure 19.1.1. However, prior to presenting the water quality application, the ideas from Sections 19.2 and 19.3 are combined for defining the intervention model of this section and explaining the model construction stages. To demonstrate that good estimates can be obtained for missing observations when there is also an external intervention, experiments are carried out with the average annual flows of the Nile River (see Figure 19.2.1 and Section 19.2.4) which were significantly lowered by the construction of the Aswan Dam. To accomplish this, in Section 19.4.4 known observations are removed from the Nile River series both before and after the intervention, and the estimates for these values are compared to the known measurements.

19.4.2 Model Description

In a qualitative fashion, an intervention model which can handle multiple interventions and missing data points can be written as

response variable = dynamic component + noise

where

dynamic component = interventions + missing data

More accurately, the above intervention model can be given as

$$y_t - \mu_y = f(\mathbf{k},\xi,t) + N_t \qquad [19.4.1]$$

where t is discrete time, y_t is the response variable which may be transformed using the Box-Cox power transformation in [3.4.30], μ_y is the mean of the entire y_t series, and N_t is the noise term which can be modelled using the ARMA model in [19.2.7]. The dynamic component, $f(\mathbf{k},\xi,t)$ contains the dynamic terms in both [19.2.1] and [19.3.1]. Consequently, $\mathbf{k}$ represents the set of transfer function parameters for modelling both the effects of the interventions and the missing data. The set ξ, contains the intervention series for modelling the occurrence and nonoccurrence of the external interventions plus the group of pulse intervention series which are needed in the intervention terms related to estimating the missing data.

When there are I_1 external interventions and I_2 missing observations in a given series, [19.2.9] and [19.3.5] can be combined to obtain

$$y_t - \mu_y = \sum_{i=1}^{I_1} \mathrm{v}_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.4.2]$$

The first summation term on the right hand side accounts for the I_1 external interventions modelled in Section 19.2.2 where $\mathrm{v}_i(B)$ has exactly the same format as the transfer function defined in [19.2.6]. For modelling the ith external intervention, the intervention series, ξ_{ti}, has a value of unity at each point in time when the intervention is taking place and values of zero elsewhere. To account for the I_2 missing data points, the second summation is designed the same way as in [19.3.5]. As explained in Section 19.3.3, the pulse intervention series, ξ_{tj}, for a missing observation, is assigned a value of one at the time of the missing data point and given values of zero elsewhere. An efficient estimate of the missing observation is the MLE of $-\omega_{0j}$.

19.4.3 Model Construction

Identification

When constructing an intervention model for handling multiple external interventions and missing observations, the appropriate tools from Sections 19.2.3 and 19.3.4 can be utilized in conjunction with the overall procedure depicted in Figure 19.2.4. Subsequent to employing exploratory data analysis tools for discovering any trends which may be caused by unknown interventions (see Section 19.2.3), an intervention model can be designed for modelling the series under consideration. Besides a sound physical understanding of the problem plus information found at the detection stage, identification procedures can be used to decide upon which parameters to include in the dynamic and noise components.

Designing the Dynamic Component: For the model in [19.4.2], a set of intervention terms are required for modelling the effects of the I_1 external interventions upon the levels of the series while another group of intervention terms are needed to estimate the I_2 missing observations. Because the format of the intervention terms for estimating the missing data is fixed, the design of these terms is considered first. As noted in Section 19.3.4, the intervention term needed for modelling the missing observation at time t_j is

$$v_j(B)\xi_{tj} = \omega_{oj}\xi_{tj}$$

where ω_{0j} is the only required transfer function parameter and ξ_{tj} is the pulse intervention series which is assigned a value of one at time t_j and zero elsewhere. Each of the intervention terms for modelling a missing observation is formulated exactly in this fashion.

From Section 19.2.3, there are two basic steps to identify each intervention term for modelling the effects of an external intervention.

1. Determine the type of changes in the time series due to each intervention. This means that a hypothesis must be made about how the series has been influenced by the intervention.
2. For each intervention, select an appropriate intervention series and associated transfer function to allow quantification of how the intervention has affected the series.

Each intervention series is usually quite simple to construct. When the external intervention is taking place, an entry in the intervention series is given a value of 1 while it is assigned a magnitude of 0 when the intervention is not occurring. The transfer function for a given intervention series must be chosen in a manner that allows the geometric shape of the dynamic response to mimic the geometrical pattern of the trend caused by the intervention in the actual series. To view the shapes of various dynamic responses for step and pulse interventions, the reader can refer to Figures 19.2.2 and 19.2.3, respectively. When dealing with seasonal data, an intervention term consisting of an intervention series and associated transfer function can be designed for each season or group of seasons that are changed in the same fashion. For instance, in Section 19.2.5 where the impacts of reservoir operation upon the average monthly flows of the S. Sask. River are modelled using intervention analysis, for the single intervention of reservoir operation, a separate intervention term is designed for each month. On the other hand, for modelling the effects of tertiary treatment upon the average monthly phosphorous levels in the Speed River, a single intervention term is used in Section 19.4.5 because all of the months are affected in a similar fashion.

A range of simple graphical techniques are available for use in step 1. When the data are seasonal, besides a plot of the entire series, it is advisable to use one or more of the following graphs for each season. Nonseasonal data can be thought of as seasonal data with only one season.

(1a) Seasonal plots.

(1b) Cusum chart (see [19.2.21] and also Figures 19.2.5 to 19.2.9).

(1c) Average plots.

(1d) Other graphs (Section 22.3).

The reader can refer to Section 19.2.3 for a detailed description of the first three identification procedures and to Section 22.3 for other useful graphs. The applications in Sections 19.2.4, 19.2.5 and 19.4.5, demonstrate how some of these graphs are used in practice.

Designing the Noise Component: Any feasible combination of the techniques outlined in Sections 19.2.3 and 19.3.4, can be employed for designing the noise term. However, a fairly straightforward procedure which should work well for most applications is the *empirical identification approach* for which related discussions appear in Sections 17.3.1, 17.5.3, 19.2.3, and 19.3.4. In particular, after identifying the form of both kinds of intervention terms required in the dynamic component, fit the model in [19.4.2] to the series where it is assumed that the noise term is white. Consequently, the intervention model has the form

$$y_t - \mu_y = \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + \sum_{i=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + a_t$$

For most applications the noise term is usually correlated. Accordingly, after obtaining the estimated residual series, $\hat{a}_t$, for the above model using the method of maximum likelihood, the kind of ARMA model to fit to the noise series can be determined by following the three stages of model construction described in Chapters 5 to 7. By using the identified form of N_t for the noise term along with the previously designed dynamic component, the intervention model in [19.4.2] is completely specified.

As an example of a specialized identification procedure which relies upon the identification tools presented in Sections 19.2.3 and 19.3.4, consider the following. Suppose there is a sufficiently long section of the series for which there are no missing values and the impacts of the external interventions are either not present or can be ignored. Simply use this part of the series to identify the parameters required in the ARMA model for N_t. Of course, when the parameters for the completely identified model are estimated, the entire series is used.

Estimation

At the estimation stage, MLE's and corresponding SE's can be simultaneously obtained for all the model parameters in [19.4.2] using the estimator described in Appendix A17.1. Of course, automatic selection criteria such as the AIC in [6.3.1] and the BIC in [6.3.5] can be employed to assist in selecting the most appropriate model by following the procedure outlined in Figure 6.3.1.

To ascertain the magnitudes of the effects of the external interventions upon the mean level of the series, the approach outlined in Section 19.2.2 can be used. Recall that for a given intervention, the change caused in the mean level of y_t is a function of the parameters in the transfer function for that intervention. Furthermore, because the SE's for the estimates of the parameters in the transfer function are obtained at the estimation stage, confidence limits can be calculated for the changes in the mean level. Practical applications for employing the formulae which describe the changes in the mean level are given in the applications of Sections 19.2.4, 19.2.5, 19.4.5, 19.5.4 and 22.4.2.

As demonstrated in [19.3.6], the MLE of the missing observation occurring at time t_j is simply $-\hat{\omega}_{oj}$. Since the SE for $-\hat{\omega}_{oj}$ is approximately normally distributed, confidence limits can be constructed for the estimated missing value. Examples of the intervention analysis approach for estimating missing data points are presented in Sections 19.3.5, 19.3.6, 19.4.4, 19.4.5 and 22.4.2.

Diagnostic Checking

In order to ascertain the adequacy of the fitted model, the residual series, $\hat{a}_t$, obtained at the estimation stage, can be subjected to stringent diagnostic checks. Tests for checking for the presence of whiteness, normality and homoscedasticity are described in Chapter 7 as well as in Sections 17.3.3, 17.5.3 and 19.2.3.

19.4.4 Experiment to Assess Data Filling when an Intervention is Present

The performance of the model in [19.4.2] for accurately estimating missing values in the presence of a known intervention is now assessed by estimating observations where the actual historical values are known. The 76 average annual flows for the Nile River at Aswan, Egypt are plotted in Figure 19.2.1. As shown graphically in this figure and more precisely by the fitted intervention model in [19.2.23], the construction of the Aswan dam in 1902 caused a significant step decrease in the mean level of the series. If the yearly data are numbered in sequential order, the intervention occurred at the thirty-third data point or $t = 33$.

The test case for testing the data filling method in the presence of an external intervention is shown in Table 19.4.1. Observations are removed before and after the intervention at the 14th and 49th data points, respectively, and replaced by values of zero at these two locations. Consequently, the intervention model consists of two pulse interventions for estimating the missing data, and, as shown in [19.2.23], one step intervention for modelling the effects of the dam upon the mean level plus a correlated noise term. Hence, the intervention model is written as

$$y_t - \bar{y} = \omega_{01}\xi_{t1} + \omega_{02}\xi_{t2} + \omega_{03}\xi_{t3} + (1 - \theta_1 B)a_t \qquad [19.4.3]$$

in which $\xi_{t1} = 1$ at $t = 14$ and $\xi_{t1} = 0$ elsewhere; $\xi_{t2} = 1$ for $t \geq 33$ and $\xi_{t2} = 0$ for $t < 33$ in order to model the intervention due to the dam; $\xi_{t3} = 1$ at $t = 49$ and $\xi_{t3} = 0$ elsewhere.

Table 19.4.1. Estimates for known observations for Nile River data.

Test Case	Lag of Missing Observation	$-\hat{\omega}_{0j}$	Standard Error	True Value, in Cubic Meters per Second
1	14	3595.38	348.73	3141.01
	49	2687.41	348.34	2377.89

From Table 19.4.1 it can be seen that the estimates for the missing data are well within two SE's of the historical values. In addition, the estimate $-\hat{\omega}_{01}$ of $-\omega_{01}$ is considerably higher than $-\hat{\omega}_{03}$. This is consistent with the drop in the mean level caused by the dam intervention for $t \geq 34$.

19.4.5 Environmental Impact Assessment of Tertiary Treatment on Average Monthly Phosphorous Levels in the Speed River

In environmental impact assessment, engineers wish to determine if a given pollution abatement procedure significantly improves the environment. Furthermore, as noted in Section 19.1, often evenly spaced environmental time series are not available, and consequently missing observations must be estimated when the impacts of the intervention are assessed. Fortunately, the flexible intervention model in [19.4.2] can easily model this type of situation.

As an interesting example, consider the graph in Figures 19.1.1 and 1.1.1 of the 72 average monthly phosphorous measurements taken downstream from the Guelph sewage treatment plant on the Speed River, Ontario, Canada. As noted in Section 19.1, in February 1974, a phosphorous removal scheme caused a significant drop in the mean level for $t \geq 26$. In addition, the filled-in circles indicate that there are three missing observations before the intervention and one missing measurement afterwards. For displaying a missing value on the graph, the missing observation is simply replaced by its monthly average across all of the months.

Notice in Figure 19.1.1 that the spread of the data is much less after the intervention date. To diminish the effects of having a smaller variance after the intervention, a natural logarithmic transformation is invoked.

Because the introduction of phosphorous treatment is expected to have an immediate effect on the water quality that persists as long as it is applied, the intervention can be modelled by a step dynamic response of the form $\omega_{04}\xi_{t4}$ in which $\xi_{t4} = 1$ for $t \geq 26$ and $\xi_{t4} = 0$ elsewhere. The four missing data points can be estimated using pulse dynamic responses as explained in Sections 19.3.3 and 19.4.2. The proposed intervention model is then written using [19.4.2] as

$$y_t - \mu_y = \omega_{01}\xi_{t1} + \omega_{02}\xi_{t2} + \omega_{03}\xi_{t3} + \omega_{04}\xi_{t4} + \omega_{05}\xi_{t5} + N_t \qquad [19.4.4]$$

in which y_t is the logarithmic transformation of the series plotted in Figure 19.1.1 and μ_y is the overall mean level of y_t; $\xi_{t1} = 1$ at $t = 6$ and $\xi_{t1} = 0$ elsewhere; $\xi_{t2} = 1$ at $t = 19$ and $\xi_{t2} = 0$ at other times; $\xi_{t3} = 1$ at $t = 25$ and $\xi_{t3} = 0$ elsewhere; $\xi_{t4} = 1$ for $t \geq 26$ and $\xi_{t4} = 0$ for $t < 26$ in order to model the phosphorous treatment intervention; and $\xi_{t5} = 1$ at $t = 41$ and $\xi_{t5} = 0$ elsewhere; and N_t is the correlated noise term.

The empirical approach is used for identifying the noise term in [19.4.4]. More specifically, the model in [19.4.4] with N_t taken to be white noise is fitted to the logarithms of the time series in Figure 19.1.1. Subsequently, an ARMA model is identified for modelling the residuals of the intervention model. Because the RACF possesses significantly large values at lower lags as well as lag 12, this indicates that nonseasonal MA parameters as well as one seasonal MA parameter may be needed in the noise term. A variety of ARMA models were considered for structuring the noise term and a suitable model was found to be a seasonal ARMA or SARMA model defined in [12.2.9] having five nonseasonal MA parameters and one seasonal MA parameters. Hence, the model for the noise term is written as

$$N_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5)(1 - \Theta_1 B^{12})a_t \qquad [19.4.5]$$

By substituting [19.4.5] into [19.4.4] the complete intervention model is revealed. Maximum likelihood estimates are then simultaneously obtained for the complete intervention model using the approach outlined in Appendix A17.1. Table 19.4.2 lists the estimates and SE's (in parentheses) for the parameters of the noise term in [19.4.5] and also the step dynamic response in [19.4.4]. As can be seen, the absolute magnitude of each of the parameter estimates is larger than twice its SE except for $\hat{\theta}_2$ which is still larger than its SE. Moreover, because the seasonal MA parameter is significantly different from zero, this confirms the importance of including this parameter in the noise term in [19.4.5].

Table 19.4.2. Parameter estimates for the phosphorous intervention model.

Parameter Estimates						
$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\Theta}_1$	$\hat{\omega}_{04}$
-0.2556 (0.1113)	-0.1467 (0.1014)	0.2870 (0.0971)	0.4657 (0.1027)	0.3303 (0.1128)	-0.3460 (0.1138)	-1.3720 (0.0720)

By substituting the estimate for ω_{04} given in Table 19.4.2 into [19.2.20], one can obtain an estimate of -74.64% for the percentage change in the mean level due to the intervention of introducing phosphorous treatment. Furthermore, by carrying out the calculations explained just after [19.2.20] in Section 19.2.2, the 95% confidence interval is found to range from -70.80% to -77.98%. Consequently, one can argue that there is a significant decrease in the phosphorous levels due to the tertiary treatment. The best estimate for this percentage drop is 74.64% while the 95% confidence interval for this decrease is from 70.80% to 77.98%. This is precisely the type of rigorous statistical statement required by environmental engineers for evaluating pollution control procedures.

It is quite interesting to note that when N_t is assumed to be white noise in [19.4.4] the estimates for the ω_{0i} parameters are significantly different than those given in Tables 19.4.2 and 19.4.3. However, when a reasonable SARMA model that is different from the one in [19.4.5] is selected for the noise term to capture correlation present in the time series, the estimates for the ω_{0i} coefficients are quite close to those listed in the two tables. This points out the practical importance of employing models, such as TFN and intervention models, for describing real world data. As mentioned in Section 17.2.4, regression analysis models do not possess the capability of handling correlated noise and hence could provide misleading results in certain situations.

Diagnostic checks indicate that based upon the available information, the model provides an adequate fit to the data. For example, the RACF for the fitted model clearly confirms the whiteness of the residuals. The Portmanteau statistic claculated using [7.3.6] for 24 lags of the RACF has a value of 17.92 on 18 degrees of freedom. Since this value is not significant at the 5% level of significance, this test also supports the whiteness assumption of the intervention model residuals.

Table 19.4.3 provides the information required for estimating the four missing values in the original phosphorous series which are listed in the bottom line of the table. More specifically, the top part of the table furnishes the negative MLE's for ω_{01}, ω_{02}, ω_{03} and ω_{05} in [19.4.4]. These four transfer function parameters link up with the observations missing at times $t = 6$, 19, 25 and 41, respectively. Below the negative of each of the parameter estimates is the mean monthly value that was inserted at the exact location in the data set where the observation was

Table 19.4.3. Estimates for the missing phosphorous data.

Parameter Estimates			
$-\hat{\omega}_{01}$	$-\hat{\omega}_{02}$	$-\hat{\omega}_{03}$	$-\hat{\omega}_{05}$
0.8304 (0.3917)	0.8169 (0.3715)	0.5429 (0.3689)	0.6479 (0.3601)
Values Used in the Data Set			
0.1524	0.2144	0.3064	0.1342
Logarithm of Values Used in the Data Set			
-1.8812	-1.5399	-1.1829	-2.0084
Estimates of Missing Values in Logarithmic Domain ($-\hat{\omega}_{0i}$ + ln of input value)			
-1.0508	-0.6780	-0.6400	-1.3605
Estimates of Missing Values in Unstransformed Domain			
0.3497	0.5076	0.5273	0.2565

unknown. As explained in Section 19.3.3, to obtain the estimate of the natural logarithm of the ith missing value, one adds $-\hat{\omega}_{0i}$ to the logarithm of the inserted value. Finally, by taking the inverse logarithmic transformation of this estimate, one obtains the estimate for each missing value in the untransformed domain as given in the bottom row of Table 19.4.3.

As demonstrated in this section, the intervention model for the Guelph phosphorous data can be employed in an environmental impact assessment study for properly determining the effectiveness of the tertiary phosphorous treatment scheme carried out at upstream sewage treatment plants. Additionally, intervention analysis can be used for estimating missing observations. Finally, the intervention model in [19.4.4] constitutes a stochastic model that can be utilized for forecasting and simulation.

19.5 INTERVENTION MODELS WITH MULTIPLE INTERVENTIONS, MISSING OBSERVATIONS AND INPUT SERIES

19.5.1 Introduction

The main objectives of this section are to describe the most *general form of the intervention model* and explain how it can be easily applied to practical problems. By combining the dynamic components from Sections 19.2, 19.3 and 17.5, a comprehensive intervention model can be defined where the dynamic component can simultaneously model the effects of multiple external interventions, estimate missing data points and describe the influence of input series upon the single output series, respectively. For example, a water quality variable, such as total organic carbon, may be the output series which can be realistically modelled using the general intervention model. Within the intervention model, it may be necessary to model the influence of a pollution abatement scheme upon the mean level of the total organic carbon series and it may be required to efficiently estimate multiple missing observations both before and after the intervention date. Furthermore, the flows in the river can be used as an input series in the model. Once again, as is the situation for the intervention models in the earlier parts of this chapter as well as the TFN models of Chapter 17, the noise term can be effectively modelled using an ARMA model.

After defining the general intervention model in the next section, the model construction stages are explained in Section 19.5.3. Although some of the material in these two sections is at least partially presented in earlier sections, for the convenience of the reader, some of the descriptions are repeated for the case of the general intervention model. In this way, practitioners who are mainly interested in the most general case of the intervention model do not have to continuously refer back to previous sections. To clearly demonstrate how an input series can be incorporated into an intervention model where the output has been influenced by an external intervention, an interesting application is presented in Section 19.5.4. An intervention model is constructed for assessing the impacts of a forest fire upon the average monthly flows of a river where average monthly riverflows from a river in a nearby basin, where there wasn't a forest fire, are used as one input series. By incorporating the input flow series into the intervention model, the effects on the output riverflows which are not due to the forest fire can be accounted for.

In Section 22.4.2 of Chapter 22, two interesting applications of intervention models containing input series are presented. In the subsection entitled the *Cabin Creek Flow Intervention Model*, an intervention model is developed for ascertaining the effects of cutting down a forest upon the average monthly flows of the Cabin Creek. Because the nearby Middle Fork River lies outside of the tree-cutting zone, it is used as an input series to remove climatic effects upon riverflows which are common to both the Cabin Creek and Middle Fork River. The intervention model also contains terms for estimating four missing values in the Cabin Creek flows and a component for modelling the impacts of the forest fire upon the Cabin Creek flows. The second application of Section 22.4.2 is described under the subsection called the *General Water Quality Intervention Model* and is concerned with designing an intervention model for determining the effects of cutting down a forest on each of a number of specified water quality variables measured in the Cabin Creek. To model the relationship between the flows and the water quality variable used in the output, the Cabin Creek Flows are used as a covariate series. In order to isolate the effects of the intervention upon the Cabin Creek water quality variable, the same water quality series from the nearby Middle Fork River where the trees were not cut down, is used as another covariate series. Finally, an intervention component is included in the model for determining the effects of clear-cutting upon the average monthly values of the Cabin Creek water quality variable.

19.5.2 Model Description

The most general format for the intervention model is written as

$$\textit{response variable} = \textit{dynamic component} + \textit{noise}$$

where

$$\textit{dynamic component} = \textit{interventions} + \textit{missing data} + \textit{inputs}$$

More precisely, the intervention model is given as

$$y_t - \mu_y = f(\mathbf{k},\xi,\mathbf{x},t) + N_t \qquad [19.5.1]$$

where t stands for discrete time, y_t is the output or response variable which may be transformed using the Box-Cox power transformation in [3.4.30], and μ_y is the theoretical mean of the entire y_t series which can be efficiently estimated using the sample mean $\bar{y}$. The noise term, N_t, accounts for the correlation in the data and can be modelled using an ARMA model. The dynamic component, $f(\mathbf{k},\xi,\mathbf{x},t)$, contains the dynamic terms from [19.2.1], [19.3.1] and also [17.5.3]. Accordingly, $\mathbf{k}$ represents the set of transfer function parameters for modelling the effects of the interventions, estimating the missing data and reflecting the influence of the input series upon the single output. The set, ξ, contains the intervention series for describing when the external interventions do and do not occur plus the group of pulse intervention series where each pulse series is assigned a value of one for the point in time for which the corresponding y_t observation is missing and is given values of zero elsewhere. As in [17.5.3] for a TFN model where there are no interventions and missing observations, the set $\mathbf{x}$ stands for the set of input series where each input series may or may not be transformed using a Box-Cox power transformation.

To fully appreciate how the general intervention model is created from the special cases discussed in previous sections, the presentation of these cases is briefly repeated here in the process of building the general form from the simpler situations.

As described in Section 19.2.2, when there are I_1 external interventions, the model at time t may be written following [19.2.9] as

$$y_t - \mu_y = \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + N_t \qquad [19.5.2]$$

where ξ_{ti} is the ith intervention series that is assigned a value of zero when the ith intervention is not in effect and given a value of unity when the ith intervention is occurring. The ith transfer function, $v_i(B)$, which is the same as the one defined in [17.5.2] for TFN models, is given as

$$v_i(B) = \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}$$

$$= \frac{\omega_{0i} - \omega_{1i}B - \omega_{2i}B^2 - \cdots - \omega_{m_i i}B^{m_i}}{1 - \delta_{1i}B - \delta_{2i}B^2 - \cdots - \delta_{r_i i}B^{r_i}}B^{b_i}$$

where $\omega_i(B)$ is the operator in the numerator of the transfer function and ω_{ji}, $j = 0,1,2,\ldots,m_i$ are the parameters of $\omega_i(B)$; $\delta_i(B)$ is the operator having the parameters δ_{ji}, $i = 1,2,\ldots,r_i$, in the denominator of $v_i(B)$ and for stability the roots of $\delta_i(B) = 0$ lie outside the unit circle; and b_i is the delay time for the ith intervention to affect y_t. The term given by $v_i(B)\xi_{ti}$, is called the *dynamic response* for the ith transfer function and ith intervention series. Plots of various kinds of dynamic responses are presented in Figures 19.2.2 and 19.2.3 for step and pulse intervention series, respectively.

As noted in Section 19.3, often there may be missing observations, especially in environmental time series. When the number of missing data points is not excessive, the intervention model can be employed for estimating the missing observations. Suppose, for example, that there are no external interventions and a time series has one missing point at time t_1. After setting the missing value y_{t_1} to zero, the intervention model for estimating the missing observation may be written following [19.3.2] as

$$y_t - \mu_y = \omega_{01}\xi_{t1} + N_t \qquad [19.5.3]$$

where ω_{01} is the parameter of the transfer function, and ξ_{t1} is the intervention series which is set to unity at time t_1 and given a value of zero elsewhere. At time t_1, [19.3.2] and [19.5.3] reduce to

$$-\omega_{01} = \mu_y + N_{t_1} \qquad [19.5.4]$$

and a maximum likelihood estimate for $-\omega_{01}$ constitutes an estimate for the missing value y_{t_1}. Because $-\omega_{01}$ depends on the noise term, N_t, the correlation structure of the series is reflected in the estimate for the missing point.

The model may be expanded to handle a situation where there is more than one missing observation. If I_2 values are missing and there are no external interventions the model in [19.3.5] is given as

$$y_t - \mu_y = \sum_{j=1}^{I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.5.5]$$

where ω_{0j} is the parameter of the jth transfer function, and ξ_{tj} is the jth intervention series which is assigned a value of unity where the jth observation is missing and zero elsewhere.

When there are I_1 external interventions and I_2 missing data points in a given series, equations [19.5.2] and [19.5.5] can be combined to obtain the result given in [19.4.2] as

$$y_t - \mu_y = \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.5.6]$$

The first summation term on the right hand side of [19.5.6] accounts for the I_1 external interventions, the second summation component allows for the I_2 missing data points, and the noise term, N_t, reflects the correlation structure of the data.

When covariate time series are available, it is possible to include them in the general intervention model. For instance, precipitation and temperature, as well as hydrologic series from nearby basins may be used as inputs for a riverflow model. For a situation where there are I_3 covariate series and no external interventions or missing data, the TFN model described in Section 17.5.2 may be written following [17.5.3] as

$$y_t - \mu_y = \sum_{k=1}^{I_3} v_k(B)(x_{tk} - \mu_{xk}) + N_t \qquad [19.5.7]$$

where x_{tk} is the kth covariate series which may be transformed using an appropriate Box-Cox power transformation, and μ_{xk} is the theoretical mean of the x_{tk} series which can be estimated using the sample mean $\bar{x}_{tk}$.

By combining [19.5.6] and [19.5.7] to form the general intervention model, it is possible to have the following comprehensive and practical model for analyzing environmental and other kinds of time series.

$$
\begin{aligned}
y_t - \mu_y = {} & \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} \\
& + \sum_{k=I_1+I_2+1}^{I_1+I_2+I_3} \nu_k(B)(x_{tk} - \mu_{xk}) + N_t \qquad [19.5.8]
\end{aligned}
$$

This comprehensive and flexible model accounts for I_1 external interventions, I_2 missing observations in y_t, and I_3 covariate series as well as reflecting the correlation structure of the series. Moreover, the model can handle both nonseasonal and seasonal data. For the case where the time series are nonseasonal, N_t can be structured using an ARMA (Chapter 3) or ARIMA (Chapter 4) model for stationary or nonstationary correlated noise, respectively. When the output and covariate time series are seasonal and follow the sinusoidal structure exhibited in Figure VI.1, they can be deseasonalized using [13.2.2] or [13.3.3] before employing the general intervention model in [19.5.8] with an ARMA noise component. Another approach is not to deseasonalize the given time series but rather permit N_t to be modelled by a SARMA or SARIMA model described in Section 12.2.1 for modelling stationary and nonstationary seasonal data, respectively. In some cases, the seasonal covariate series in [19.5.8] may remove all or part of the seasonality contained in the response series and thereby cause the noise to be nonseasonal or else slightly seasonal. Finally, in addition to environmental impact assessment, data filling, and causality modelling, the finite difference equation model in [19.5.8] can be utilized for forecasting and simulation.

19.5.3 Model Construction

When developing a general intervention model to fit to a set of time series, a sound physical understanding of the problem in conjunction with the overall procedure outlined in Figure 19.2.4 can be used. In order to detect trends in a series which may be caused by unknown interventions, the exploratory data analysis tools which are briefly referred to in Section 19.2.3 and described in detail in Section 22.3 of Chapter 22, can be utilized. Additionally, the nonparametric trend tests of Section 23.3 and robust locally weight regression smooth of Section 24.2.2 can be employed for discovering unknown trends and confirming the presence of suspected trends caused by known external interventions. After finding suitable physical explanations to account for all of the external interventions, the parameters required in the intervention model must be decided upon. For all of the special cases of the intervention models presented in this chapter, the main differences in constructing the models occur at the identification stage. Consequently, the discussion in this section concentrates on model identification. Following model identification, MLE's can be obtained for the model parameters and the adequacy of the fitted model can be checked.

Identification

The instructions for identifying the special cases of the intervention model are given in the following sections:

(1) Section 19.2.3 for the intervention model in [19.2.9] and [19.5.2] which can handle multiple interventions.

(2) Section 19.3.4 for the intervention model in [19.3.5] and [19.5.5] which can be used for estimating missing observations.

(3) Section 19.4.3 for the intervention model in [19.4.2] and [19.5.6] that can model the effects of multiple external interventions upon the mean level of y_t and also be used for estimating missing observations.

(4) Sections 17.5.3 and also 17.3.1 for the TFN model in [17.5.3] and [19.5.7] which can handle multiple input series.

In order to design the general intervention model in [19.5.8], appropriate identification tools from all of the foregoing sections must be selected. This means that there are quite a few different approaches which could be adopted. The most convenient procedures for identifying the general intervention model are now discussed separately for the dynamic and noise components.

Designing the Dynamic Component: For the general intervention model in [19.5.8], three distinct kinds of terms are needed in the dynamic component. A set of intervention terms are required to model the effects of the I_1 interventions, a group of intervention terms are needed to estimate the I_2 missing observations, and a set of dynamic responses are required for describing how the I_3 inputs influence the single output. When designing the dynamic component, it is most convenient to separately design the three parts of the dynamic component.

Missing values. Because the form of each intervention term needed for modelling a missing observation is fixed, the design of the terms for modelling the missing observations is entertained first. As explained in Section 19.3.4 and as shown in [19.5.3] and [19.5.5], the intervention term needed for modelling the missing observation at time t_j is

$$v_j(B)\xi_{tj} = \omega_{0j}\xi_{tj}$$

where ω_{0j} is the only required transfer function parameter and ξ_{tj} is the pulse intervention series which is assigned a value of unity at time t_j and zero elsewhere. Each of the intervention terms for modelling a given missing data point is written in exactly the same manner. The MLE of $-\omega_{0j}$ constitutes an efficient estimate for the missing value at time t_j.

The reader should bear in mind that the general intervention model in [19.5.8] can only be utilized for estimating the missing observations in the output series y_t. If there are missing observations in an input series, x_{tk}, the model in [19.5.5] can be employed to estimate the missing observations where x_{tk} and μ_{tk} replace y_t and μ_y, respectively, in [19.5.5]. Subsequent to estimating all of the missing measurements separately for each x_{tk} series, the input series can be employed in the overall intervention model in [19.5.8].

External interventions. As described in Section 19.2.3, there are two basic steps for identifying each intervention term needed for modelling the impacts of an external intervention upon the mean level of y_t.

(1) Determine the type of change in the time series due to each intervention. Hence, a hypothesis must be made on how the y_t series has been altered by the intervention.

(2) For each intervention, choose an appropriate intervention series and associated transfer function to permit quantification of how the intervention has influenced the y_t series.

Usually, each intervention series can be easily designed. Whenever the external intervention is occurring, the entries are assigned values of one while they are given zero values when the intervention is not taking place. For a given intervention series, the transfer function must be designed in a manner that permits the geometric shape of the dynamic response to mimic the geometrical pattern of the trend caused by the intervention in the y_t series. Graphs of various dynamic responses caused by step and pulse interventions are displayed in Figures 19.2.2 and 19.2.3, respectively. When dealing with seasonal data, an intervention term consisting of an intervention series and associated transfer function can be identified for each season or groups of seasons that are changed in the same fashion.

For employment in step 1, a variety of informative, yet simple, graphical techniques are available. When considering seasonal data, in addition to a plot of the y_t time series against time, one or more of the following graphs can be drawn for each season. Of course, nonseasonal data can be thought of as seasonal data with only one season per year.

(1a) Seasonal plots.

(1b) Cusum chart (see [19.2.21] and also Figures 19.2.5 to 19.2.9).

(1c) Average plots.

(1d) Other graphs (Section 22.3).

The reader can refer to Section 19.2.3 for a detailed description of each of the first three identification graphs and to Section 22.3 for other useful graphs. The applications in Sections 19.2.4, 19.2.5 and 19.4.5 illustrate how some of these graphs are used in practice.

Inputs. In [17.2.5], a TFN model is defined where there is only one input series x_t which affects the output series y_t. The transfer function which describes how the x_t series affects the output can be designed by using one or more of the following identification techniques which are described in detail in Section 17.3.1.

(1) Empirical identification approach

(2) Haugh and Box identification method

(3) Box and Jenkins identification procedure.

The application presented in Section 17.4.2 shows how each of the above techniques can be used in practice.

As noted in Sections 17.3.1 and 17.5.3, all three identification methods were developed under the assumption that there is only one input series present in the model and the input series only affects the output. When there is more than one input series, the obvious way to use each identification procedure, especially the second and third ones, is to investigate, pairwise, the relationship between each x_{tk} series and y_t in order to design the form of the transfer function $v_k(B)$. Nevertheless, in a general intervention model with more than one covariate series, the covariate series may affect one another besides influencing the response variable y_t. When there is not too much interaction among the I_3 input series, fairly correct transfer functions may be identified using the pairwise identification procedure. Whatever the case, the assumptions that the x_{tk}'s are independent is not assumed in the general intervention model in [19.5.8] and the TFN model in [17.5.3]. Consequently, if required, a number of tentative dynamic models for the input series can be considered when estimating the parameters for the resulting overall general intervention models, where, of course, tentative designs for the noise component are assumed. A discrimination technique such as the AIC in [6.3.1] can then be utilized to choose the most appropriate general intervention model.

Probably, the simplest approach for designing the I_3 transfer functions, especially when there are more than two input series, is to employ the empirical approach. If there is difficulty in designing one or more of the transfer functions, one or both of the other two identification methods can be used in conjunction with the empirical approach. The reader should keep in mind that if the Haugh and Box or Box and Jenkins approach is used, the effects of the interventions upon y_t must somehow be removed or accounted for before calculating the required CCF's (cross-correlation functions). For instance, suppose there is a sufficiently long portion of data for which the impacts of the interventions upon y_t can be neglected or else are not present and there are no missing values. Then this section of the data can be used to calculate the CCF's needed in the two approaches. Another method is to first fit the intervention model in [19.5.6] to the y_t series where the I_3 input series are not included in the model. Consequently, from [19.5.6]

$$N_t = (y_t - \mu_y) - \left\{ \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} \right\}$$

The estimated noise series, $\hat{N}_t$, in [19.5.6] can be thought of as an estimate of the y_t series where the I_1 missing values have been estimated and the effects of the I_2 interventions have been removed. Note that N_t series can be estimated even prior to designing the ARMA model to describe N_t. Simply assume that N_t is white in [19.5.6] and a program can be used to estimate the residual series which will probably be correlated. This correlated residual series constitutes the estimate for N_t. Using the $\hat{N}_t$ series, the necessary CCF's needed in the Haugh and Box, and the Box and Jenkins methods can be calculated for the entire series following the detailed procedures outlined in Section 17.3.1.

The authors have found in practice that usually the *empirical approach* works well for designing the transfer functions needed for the I_3 input series and, therefore, it is usually not necessary to obtain the N_t series described in the previous paragraph. As explained in Sections 17.3.1 and 17.5.3, the empirical approach is straightforward to use but it does require the modeller to exercise good judgement. Based upon an understanding of the physical phenomena that generated the y_t and x_{tk} time series as well as the mathematical properties of the general intervention model, each transfer function, $v_k(B)$ can be identified. For example, suppose that the output is an average monthly series such as total organic carbon and that one of the input series is precipitation. It may be known from the physical characteristics of the watershed that rainfall for the current month only affects the total organic carbon for that month. Consequently, to model the precipitation series, x_{tk}, it may be appropriate to employ the transfer function

$$v_k(B) = \omega_{0k}$$

A water quality application where a transfer function like this is employed is presented in Section 22.4.2.

Designing the Noise Component: The best procedure for identifying the noise component is to employ the *empirical approach* for which earlier related discussions appear in Sections 17.3.1, 17.5.3, 19.2.3 and 19.4.3. After identifying the form of the complete dynamic component, fit the model in [19.5.8] to the series where it is assumed that the noise term is white. Hence, the general intervention model has the form

$$y_t - \mu_y = \sum_{i=1}^{I_1} v_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + \sum_{k=I_1+I_2+1}^{I_1+I_2+I_3} v_k(B)(x_{tk} - \mu_{tk}) + a_t$$

For most applications, the noise term is correlated. Therefore, after obtaining the estimated residual series, $\hat{a}_t$, for the above model using the method of maximum likelihood, the type of ARMA model to fit to the noise series can be determined by following the three stages of model construction described in Chapters 5 to 7. The identified noise term along with the previously designed dynamic component, provides the complete design for the intervention model in [19.5.8].

Estimation

The MLE's and SE's for all of the parameters in the general intervention model are simultaneously obtained at the estimation stage using the estimator described in Appendix A17.1. When there are a range of tentative models to choose from, automatic selection criteria such as the AIC in [6.3.1] and the BIC in [6.3.5] can be employed for discrimination purposes by following the general procedure of Figure 6.3.1.

For calculating the affects of the external interventions upon the mean level of the y_t series, the approach described in Section 19.2.2 can be utilized. For a given intervention, the change caused in the mean level of y_t is a function of the parameters in the transfer function used with the corresponding intervention series. By considering the standard errors of estimation for the transfer function parameters, confidence limits can be obtained for the changes in the mean level.

As explained in [19.3.6], the MLE of the missing observation at time t_j, is given by $-\hat{\omega}_{0j}$. By considering the SE for $-\hat{\omega}_{0j}$, confidence limits can be obtained for the estimate of the missing value.

Diagnostic Checks

At the estimation stage, the residual series, $\hat{a}_t$, is estimated. To test the adequacy of the fitted model, these residuals can be subjected to diagnostic checks. All the diagnostic checks for the residual series presented in Chapter 7 and elsewhere can be used for checking the whiteness, normal, and homoscedastic assumptions of the residuals. To verify that the residuals are white, the recommended procedure is to plot the RACF in [7.3.1] along with appropriately chosen confidence limits. Additionally, the cumulative periodogram in [2.6.2] and the modified Portmanteau test in [17.3.7] can be used to determine whether or not the residuals are uncorrelated. When the residuals are correlated, the model is inadequate and appropriate changes must be made to the model by repeating the stages of model development in Figure 19.2.4. As is the case with most of the models discussed in this book, if the residuals do not follow a normal distribution and/or are heteroscedastic, an appropriate Box-Cox transformation of the y_t series and perhaps also some of the x_{tk} series may rectify the situation.

In Sections 17.3.3 and 17.5.3, additional tests are given for TFN models where there are single or multiple input series, respectively. As noted in Section 17.3.3, if the residual ACF indicates that the residuals are correlated, the model inadequacy could be due to the noise term, the transfer functions in the dynamic component, or both. The form of the significant autocorrelations present in the estimated residual ACF may indicate what type of model modifications should be made. Additionally, assuming that the transfer functions and intervention series for modelling the interventions are correctly designed, investigation of the form of the CCF between each prewhitened x_{tk} series and $\hat{a}_t$ may also assist in detecting where the sources of the problems are located and how they should be rectified.

Fortunately, in practice the authors have never found it necessary to locate errors in a general intervention model by investigating the relationship between a prewhitened x_{tk} series and $\hat{a}_t$. Usually, any problems with the design of the model can be detected and eliminated by simply examining the RACF and repeating the appropriate stages of model construction.

19.5.4 Effects of a Forest Fire upon the Spring Flows of the Pipers Hole River

Case Study

An intervention model is developed for modelling the effects of a natural intervention upon the mean level of an average monthly hydrological time series. In particular, an intervention model is determined to describe the consequences of a forest fire on the spring flows of the Pipers Hole River in Newfoundland, Canada. As is shown, the model is capable of explaining how the spring flows gradually recover their previous stochastic characteristics before the forest fire as the new forest slowly grows over the years. The intervention model also contains an input series, which is an average monthly riverflow series at a nearby river basin where there was no forest fire. Even though there was a large forest fire, the series does not contain any missing values. Consequently, the only part of the dynamic component in the general intervention model in [19.5.8] which is not included in the intervention model for the Pipers Hole River, is a set of terms for estimating missing observations. Earlier presentations of this application are given by Hipel et al. (1977b) and Hipel et al. (1978). For a water quality application of intervention analysis where there are two input series, a single intervention, plus missing data, the reader can refer to Section 22.4.2.

The Pipers Hole River is located in the southeastern part of the province of Newfoundland in Canada and covers an area of 829 km^2. The drainage area consists of 88 km^2 of lakes, 176 km^2 of bog, 461 km^2 of barrens and 104 km^2 of forest. The basin is uninhabited and there is no access road to the interior.

The Pipers Hole River drains the basin into the head of Placentia Bay, which forms part of the Atlantic Ocean along the coast of Newfoundland. A gauging station located near the mouth of the river has been in continuous operation since 1953 and records the natural runoff from 777 km^2 of the drainage area.

During the period from August to October of 1961, a major fire destroyed an expanse that included 85% of the Pipers Hole drainage basin. In addition to some fir and various deciduous species, the major tree type in the basin prior to the fire was spruce. The fire devastated most of the forest and all other forms of vegetation that were within its path. The shallow soil mantle in the lower reaches of the basin was incinerated and consequently surface boulder was exposed over most of the area.

A unique application of intervention analysis is to develop a stochastic model for the monthly flows of the Pipers Hole River that incorporates the effect of the forest fire intervention on the riverflows. A forest fire can have transitional impacts on riverflows that must be included in an intervention model. Because the surroundings are denuded of all vegetation, this causes initial sudden changes in the flow regime of a river. However, over the years as the vegetation recovers, the riverflows gradually revert to their previous state.

The Bay du Nord River is located 69 km west of the Pipers Hole River and was untouched by the 1961 fire. Flow records have been tabulated continuously since 1952 and at the location of the measuring gauge, the Bay du Nord River drains an area of 1176 km^2. Because of their geographic proximity, these two basins have identical climates and the Bay du Nord basin possesses a vegetation cover that is similar to that of the Pipers Hole River vicinity prior to the fire. Therefore, the Bay du Nord flows are suitable for comparison to those of the Pipers Hole River. By including the Bay du Nord flows in the intervention model for the Pipers Hole River, flow changes that are not due to the forest fire but are a result of climatic conditions are automatically accounted for. In this way, the intervention component of the model only describes changes resulting from the fire.

Model Development

Identification: Qualitatively, an intervention model for the Pipers Hole River can be written as

Pipers Hole flows = dynamic component + noise

where

dynamic component = fire intervention + Bay du Nord flows

To identify the dynamic and noise components, the empirical approach of Section 19.5.3 is employed.

Large riverflows in Newfoundland occur in the spring due to snow melt. Consequently, when considering average monthly flows, a forest fire may cause significant alterations in flow patterns during the spring months. An inspection of separate monthly plots from January 1953 to December 1973 reveals that the flows for March and April may be changed by the fire. The flows in these months appear to increase immediately after the fire, followed by a steady decrease to former levels over the years. Because this type of variation does not occur in the Bay du Nord monthly riverflows, this suggests that the changes in the Pipers Hole River flows, excluding intrinsic random variation, are due solely to the forest fire.

Considering the aforesaid facts, a tentative design for the intervention component is

$$\textit{intervention component} = \frac{\omega_{01}}{(1-\delta_{11}B^{12})}\xi_t \qquad [19.5.9]$$

where

$$\xi_t = \begin{cases} 1, & t = \text{March 1962, April 1962} \\ 0, & \text{otherwise} \end{cases}$$

is the intervention time series.

The ω_{01} parameter represents the initial change in the March and April flows due to the fire. The denominator of the transfer function models the gradual return of the spring flows to previous levels due to vegetation regeneration. This effect is more easily visualized by expanding the dynamic response for the intervention as

$$\frac{\omega_{01}}{(1-\delta_{11}B^{12})}\xi_t = \omega_{01}(1+\delta_{11}B^{12}+\delta_{11}^2B^{24}+\delta_{11}^3B^{36}+\cdots)\xi_t \qquad [19.5.10]$$

Because $|\delta_{11}| < 1$, the infinite series expansion in [19.5.10] is convergent and events further into the past have a decreasing influence on the present. The ξ_t series is zero before the intervention so that [19.5.10] is only non-zero for the months of March and April after 1961. As the years progress subsequent to the fire, the value of the dynamic response in [19.5.10] for these two months decreases asymptotically to zero.

For seasonal riverflow data, it has been found in practice that taking natural logarithms of the data is a reasonable transformation to remove heteroscedasticity and non-normality of the residuals. A possible intervention model for the forest fire problem is

$$y_t - \mu_y = \frac{\omega_{01}}{(1-\delta_{11}B^{12})}\xi_t + \omega_{02}(x_t - \mu_x) + N_t \qquad [19.5.11]$$

where y_t is the series of natural logarithms of the average monthly Pipers Hole Riverflows, μ_y is the mean of the entire y_t series, x_t is the sequence of natural logarithms of the Bay du Nord Riverflows, and μ_x is the mean of the x_t series. Because of similar climatic conditions, the ω_{02} parameter reflects the fact that for each month the flow in the Bay du Nord River behaves similar to that in the Pipers Hole River. In other words, the dynamic response in [19.5.11], due to the Bay du Nord flows, models the portions of the Pipers Hole River data that are common to both rivers.

The empirical approach to identify the form of the noise term is to initially assume that N_t is white so that [19.5.11] becomes

$$a_t = (y_t - \mu_y) - \left[\frac{\omega_{01}}{1-\delta_{11}B^{12}}\xi_t + \omega_{02}(x_t - \mu_x)\right]$$

Subsequent to obtaining the estimated residual series, $\hat{a}_t$, for the above model by simultaneously estimating all the model parameters using the method of maximum likelihood, the type of ARIMA model to fit to $\hat{a}_t$ can be identified. Because the ACF of $\hat{a}_t$ has values which are significantly different from zero at lags 1 and 12, this suggests that $\hat{a}_t$ and hence N_t can be modelled by a seasonal ARIMA $(0,0,1)(0,0,1)_{12}$ process from [12.2.9] as

$$N_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12})a_t \qquad [19.5.12]$$

Notice that neither seasonal or nonseasonal differencing are required. This is because the covariate series, x_t, in [19.5.11] causes the nonstationary part of the seasonality to be removed from the response, y_t. Consequently, for this application, the inclusion of a covariate series in the intervention model eliminates the need for differencing or deasonalizing the y_t series, thereby decreasing the number of parameters required in the overall intervention model.

Estimation: By incorporating the design of N_t given by [19.5.12] into [19.5.11], the intervention model for the Pipers Hole River is completely specified as

$$y_t - \bar{y} = \frac{\omega_{01}}{(1 - \delta_{11}B^{12})}\xi_t + \omega_{02}(x_t - \bar{x}) + (1 - \theta_1 B)(1 - \Theta_1 B^{12})a_t \qquad [19.5.13]$$

In Table 19.5.1, the MLE's and SE's for the parameters in the above model are listed.

Table 19.5.1. Forest fire intervention model parameter estimates.

Parameter	Estimate	Standard Error
ω_{01}	0.392	0.200
δ_{11}	0.946	0.091
ω_{02}	1.201	0.047
θ_1	-0.228	0.059
Θ_1	-0.143	0.068

Model Adequacy: A range of diagnostic checks are executed to insure that the $\hat{a}_t$'s are independent, homoscedastic and normally distributed. In all cases, the tests reveal that the general intervention model in [19.5.13] adequately models the data. For example, the portmanteau statistic Q_L in [7.3.6] has a value of 25.62 for 35 degrees of freedom. This indicates that based on the available data, the $\hat{a}_t$'s are independent because this value is not significant even at the 50% level of significance. From Section 7.5.2, the statistic used to test for changes in the variance of the residuals, depending on the current level of the series, has a value of 7.729, while the statistic for variance changes, depending on time, has a value of 0.159. The former is not significant at the 0.5% significance level, while the latter is not significant at the 50% level. The residuals possess no significant skewness because g in [7.4.1] has a value of -0.0520 with a SE error of 0.1936.

Effects of the Forest Fire

The general procedure outlined in Section 19.2.2 can be used to ascertain how the forest fire has affected the mean level of the spring flows of the Pipers Hole River. This is effected by taking antilogarithms and expected values of [19.5.13] before and after the intervention.

Because natural logarithms were taken of the riverflows in [19.5.13], to express the intervention effects in terms of the Pipers Hole Riverflows, a transformation must be calculated. Taking the natural antilogarithms of [19.5.13] gives

$$Y_t = \left(e^{\bar{y}}e^{-\omega_{02}\bar{x}}\right)\left(e^{\omega_{02}x_t}e^{N_t}\right)e^{\frac{\omega_{01}}{1-\delta_{11}B^{12}}\xi_t}$$

$$= c'_1\left(e^{\omega_{02}x_t}e^{N_t}\right)e^{\frac{\omega_{01}}{1-\delta_{11}B^{12}}\xi_t} \qquad [19.5.14]$$

where

$c'_1 = e^{\bar{y}}e^{-\omega_{02}\bar{x}}$, a constant.

Before the intervention, ξ_t has a value of zero and therefore taking expectations of [19.5.8] produces

$$E[Y_t]_{\text{before}} = c'_1c'_2 \qquad [19.5.15]$$

where

$$c'_2 = E[e^{\omega_{02}x_t}e^{N_t}]$$

After the fire, ξ_t has a value of unity for March and April of 1962 and is zero at all other times. The expected value of Y_t in [19.5.14] for each year after the fire in 1961 is

$$E[Y_t]_{\text{after}} = c'_1c'_2e^{\omega_{01}\delta_{11}^{(\text{date}-1962)}} \qquad [19.5.16]$$

where

date stands for any year after 1961.

Using [19.5.15] and [19.5.16], the percentage increase in the spring runoff in March and April for any year after the fire is

$$\%\text{ increase} = \left(\frac{E[Y_t]_{\text{after}}}{E[Y_t]_{\text{before}}} - 1\right)100$$

$$= \left(e^{\omega_{01}\delta_{11}^{(\text{date}-1962)}} - 1\right)100 \qquad [19.5.17]$$

where the MLE's of ω_{01} and δ_{11} are listed in Table 19.5.1.

By utilizing [19.5.17] the percentage increase in the spring runoff can be calculated for each year after the fire. Table 19.5.2 shows that as the vegetation continues to mature after the fire the percentage increase in flow will subside over the years and by the year 2000 it should be only about 4.5% greater than it was before the fire. This argument is of course valid only if the Pipers Hole River basin is not subject to any other major natural or man-induced interventions in the interim.

Table 19.5.2. Percentage increase in spring runoff after the fire.

Date	% Increase in Spring Runoff
1962	47.95
3	44.83
4	41.93
5	39.25
6	36.76
7	34.44
8	32.29
9	30.29
1970	28.42
1	26.68
2	25.06
3	23.55
4	22.13
5	20.81
6	19.57
7	18.41
8	17.32
9	16.31
1980	15.35
1990	8.50
2000	4.50

19.6 PERIODIC INTERVENTION MODELS

19.6.1 Introduction

As emphasized by authors such as Moss and Bryson (1974), seasonal hydrological and other types of time series exhibit an autocorrelation structure which depends on not only the time lag between observations but also the season of the year. Furthermore, within a given season, usually second order stationarity is preserved by natural time series. For example, at a location in the northern hemisphere the monthly temperature for January across the years may fluctuate with constant variance around an overall mean of $-5\ ^oC$. In addition, the manner in which the January temperature is correlated with December and November as well as the previous January may tend to remain the same over the years. To model this type of series, which possesses seasonal sinusoidal characteristics similar to the seasonal hydrological time series shown in Figure VI.1, one can employ the periodic models described in Chapter 14. In particular, the PAR (periodic autoregressive) model is defined in [14.2.1], by fitting a separate AR model to each season of the year. As shown in [14.2.15], a PARMA (periodic ARMA) model can also be used to model seasonal time series by having a separate ARMA model for each season of the year.

A natural extension of the periodic models of Chapter 14, is to define periodic intervention models and TFN models. In particular, to obtain a periodic intervention model for the most general situation shown in [19.5.8], a suitable subscript can be added to each parameter and series to indicate that a separate intervention model is fitted to each season of the year. When there are no interventions or missing data, the periodic intervention model would become the periodic TFN model which in turn is the periodic version of the TFN model in [17.5.3].

To fit a periodic intervention model to a given set of data, the modelling stages of Figure 19.2.4 can be followed. In general, most of the construction tools of Chapter 19 can be used with periodic intervention models, where appropriate modifications are made whenever necessary. Subsequent to identifying which parameters to include in the intervention model for each season of the year, the method of maximum likelihood can be utilized to obtain efficient estimates of the model parameters. The estimated model residuals can then be subjected to the diagnostic tests described in Section 14.3.4 for the residuals of the PAR models.

A drawback of the periodic intervention model is that it requires many more parameters than its nonperiodic counterpart. To reduce the number of parameters, only those terms of the model which are required to be periodic can be defined in a periodic manner. In fact, this approach is already used in a previous application in Section 19.2.5 of this chapter. In that section, an intervention model is developed for modelling the effects of reservoir operation upon the mean level of the average monthly flows of the S. Sask. (South Saskatchewan) River. Notice in [19.2.24] that there is a separate intervention term for each month or season of the year and hence the dynamic component is designed to be periodic. However, in [19.2.24] the noise term is not periodic since there is only one noise term for use across all the months. To have a completely periodic model for the S. Sask. flows there would have to be a separate intervention and noise component for each season of the year. A periodic intervention model for the S. Sask. River is developed in the next subsection.

19.6.2 Periodic Intervention Model for the Average Monthly Flows of the South Saskatchewan River

Recall from Section 19.2.5 and also from Figure 19.2.11, that the Gardiner dam on the S. Sask. River came into operation in January, 1969. To define a periodic intervention model for modelling the average monthly flows of the S. Sask. River, consider the situation given by Hipel and McLeod (1981) where the noise term is AR(2) for each season or month of the year. Then for the mth month the periodic intervention model is given by

$$y_{r,m} - \mu_m = \omega_{0m}\xi_{rm} + \frac{a_{r,m}}{1 - \phi_{1,m}B - \phi_{2,m}B^2} \qquad [19.6.1]$$

where $y_{r,m}$ stands for the response series consisting of the S. Sask. flows in the rth year and mth month where for this application the response series is first transformed by taking natural logarithms, μ_m is the mean of $y_{r,m}$ for the mth month, and $a_{r,m}$ is the innovation sequence for the rth year and mth month. For convenience, the ith previous value to $y_{r,m}$ can be denoted by $y_{r,m-i}$ for $i = 1,2,\cdots$, so that, for example, $y_{9,12}$, $y_{10,0}$ and $y_{8,24}$ all refer to the same observation for monthly data where the number of seasons is 12. The intervention parameter ω_{0m} is used to reflect the impact of reservoir operation upon the mth season for which the intervention series ξ_{tm} is assigned a value of zero before 1969 and a value of one from 1969 onwards. The periodic noise term in [19.6.1] is a special case of the PAR model in [14.2.1] where the AR operator for the mth season is of order two and has the parameters $\phi_{1,m}$ and $\phi_{2,m}$.

Let the mean for the mth season for the PAR model in [14.2.1] be denoted as μ'_m. Then, by allowing μ'_m to be represented by

$$\mu'_m = \mu_m + \omega_{om}\xi_{tm} \qquad [19.6.2]$$

the same estimation procedures used with the PAR models can be employed for estimating the parameters of the model in [19.6.1] for each season of the year. Following the approach used to derive [19.2.25], the intervention parameter for each month or season can be converted to the percentage change in the mean level for that month by using

$$\% \text{ change} = (e^{\omega_{0m}} - 1)100 \qquad [19.6.3]$$

After estimating all the model parameters in [19.6.1] for each month of year, the estimated values for each ω_{0m}, $m = 1,2, \cdots ,12$, are substituted into [19.6.3] to obtain the percentage change in the mean level for each month. Table 19.6.1 lists the estimated percentage change in the mean level for each month during the period from 1969 to 1974. Notice that these results are similar to those given in Table 19.2.4 where the quasi-periodic intervention model in [19.2.24] is used to model the S. Sask. River flows. Consequently, for this application the model in [19.2.24] probably possesses enough complexity to adequately model the data. However, in other situations it may be necessary to use a completely periodic intervention model as is done in [19.6.1].

Table 19.6.1. Estimated percentage changes in the average monthly flows of the S. Sask. River at Saskatoon from 1969 to 1974.

Month	Percentage Change	Month	Percentage Change
January	450.09	July	-53.23
February	405.84	August	-28.26
March	180.34	September	-10.90
April	-40.34	October	35.22
May	-52.26	November	123.45
June	-63.91	December	339.85

19.6.3 Other Types of Periodic Intervention Models

When deemed necessary, appropriate adjustments can be made to the periodic model to make it either simpler or more complex. Because a simpler form of the periodic model is discussed with the S. Sask. application in Section 19.2.5, consider the case where the complexity of the periodic intervention model must be increased. For instance, suppose it is suspected that the noise term may be affected by an intervention. Then for each season of the year there would be a separate noise term for both before and after a given intervention. In fact, to allow all of the parameters in a periodic model to change as time progresses, the model could be defined within the Kalman filtering approach to modelling. Whatever the case, a given model should only possess a level of complexity which is just high enough to allow the fitted model to adequately model the data under consideration. In this way, there will be just enough parameters to provide a good statistical fit to the data where the overall format of the model provides a suitable range of intervention models to be entertained.

19.7 DATA COLLECTION

In Section 1.2.3, it is pointed out that a *scientific investigation* involves the following two main tasks (Box, 1974):

1. the *design problem* for which the appropriate data to obtain at each stage of an investigation must be decided upon.
2. the *analysis problem* where models are employed for determining what the data entitles the investigator to believe at each stage of the investigation.

In the previous sections of this chapter, the analysis problem is mainly entertained by fitting intervention models to time series in order to ascertain whether or not interventions caused significant changes in the mean levels of the series. Consequently, within this section some comments are made about the design or data collection problems.

When dealing with time series studies, often the data were collected over a long period of time and the professionals analyzing the collected data did not take part in designing the data collection procedure in the first place. For example, for the data considered in the applications in this book, the authors had to rely upon data which were already collected by various agencies. Nevertheless, practitioners are advised wherever possible to actively take part in the design of the scheme for collecting the data which they will analyze.

Even though the authors were not involved in the design of the data collection schemes for the data used in this book, they still have control over which of the collected data to use. For instance, for the applications of Sections 19.5.4, 22.4.2, 17.4.2, 17.4.3 and 17.5.4, various covariate series can be incorporated into the intervention or TFN models. By appropriately selecting which covariate series to include in the models, the authors take full advantage of the data bases which are available. In all of the aforesaid applications, the consideration of suitable input series makes the ensuing analyses much more accurate.

For specialized types of intervention models, Lettenmaier et al. (1978) clearly show how the design of the data collection scheme is directly related to the form of the intervention model which will eventually be used to analyze the collected time series. In other words, the design and analysis problems are interrelated with one another. By having a knowledge of what type of analytical tools will eventually be used to extract and interpret information from the data, an optimal data collection scheme can be designed. Consequently, whenever possible, scientists should be involved with both the design and analysis activities for a given investigation.

Based upon a knowledge of the *variance-covariance matrix* for a given intervention model (see Appendix A6.2 for a discussion of the information and variance-covariance matrices), Lettenmaier et al. (1978) derive a power function for that model. The power is considered to be the probability of detecting the existence of an intervention response function when one is actually present. The power function can easily be shown to be a function of a number of factors which include the number of variables in the intervention model, the sample size and the number of observations before and after the intervention. By investigating the properties of power functions for a number of specific intervention models, Lettenmaier et al. (1978) come up with a number of suggestions for data collection which include:

1. As is also pointed out by Lettenmaier (1978), data should be collected using a uniform sampling frequency. This is because the intervention model, as well as the other time series models in this book, are defined under the assumption that the data are evenly spaced

over time.

2. If demands from multiple users require nonuniform sampling frequencies, then the data collection scheme should be designed to allow efficient estimates to be obtained for a time series where the data points are equally spaced over time (also see the discussion in Section 19.3.2 for filling in missing data).
3. As would be expected, uniformly spaced data are required both before and after the date of intervention in order to calibrate the intervention model.
4. Intuitively, one may think that equal amounts of data should be collected both before and after the intervention. However, for three of the four specific intervention models considered by Lettenmaier et al. (1978), it is advantageous to have a longer record after the intervention takes place. This could be due to the fact that an intervention term only appears in the intervention model after the intervention is in effect (recall that the intervention series is assigned values of zero before the intervention date).
5. The threshold (minimum) level of change that can be detected is quite high unless sample sizes of at least 50 and preferably 100 are available.
6. The threshold level is dependent upon the complexity of the intervention model and, as would be anticipated, more complex models require larger sample sizes.

19.8 CONCLUSIONS

As demonstrated by the wide range of applications in this chapter and also Section 22.4.2, intervention analysis constitutes a flexible and comprehensive approach for realistically modelling many types of situations which can arise in practice. The efficacy of the intervention model for realistically modelling many kinds of practical problems can be directly attributed to its clever *mathematical design*. Qualitatively, an intervention model can be written as

response variable = dynamic component + noise

For all of the special cases of the intervention model which are discussed in the book, it is assumed that there is a single output or response variable and that the noise term can be described by an ARMA or ARIMA model. However, the different types of dynamic components which can be incorporated into the overall intervention model are as follows:

1. To model the effects of one or more man-induced and/or natural interventions upon the mean level of the output, in Section 19.2 the dynamic component is simply given as

   dynamic component = interventions

2. If there are missing data in a series, the procedure of Section 19.3 can be used where

   dynamic component = missing data

3. When in addition to missing data the output series is acted upon by one or more external interventions, in Section 19.4 the dynamic component is defined as

   dynamic component = interventions + missing data

4. When the single output series is affected by one or more input or covariate series and there are no interventions or missing data, the intervention model is the same as the TFN model of Chapter 17. In fact, as noted in Section 19.1, the intervention model can be considered

as a special kind of TFN model for which appropriate designs are incorporated into the dynamic component to model the effects of the interventions and estimate the missing data. When there are only multiple input series, the dynamic component from Section 17.5 is given as

dynamic component = inputs

5. In Section 19.5, the dynamic component is defined to handle all of the foregoing situations such that

dynamic component = interventions + missing data + inputs

The realistic mathematical design of the intervention model constitutes a "necessary condition" for the model to be useful for properly studying actual time series. To achieve the "necessary and sufficient conditions" for successful modelling, flexible *model construction* tools are needed in order to decide upon which parameters are required in the intervention model for modelling a given data set. Combined with a thorough physical understanding of the problem being investigated, these model construction tools can be used within the overall framework of model construction stages portrayed in Figure 19.2.4. As described in Sections 19.2.3 and 22.3, *exploratory data analysis* tools can be employed for detecting the effects of any unknown interventions. Subsequent to this, identification techniques can be used for deciding upon which parameters to include in the dynamic and noise components. A wide variety of *identification methods* are described in Sections 19.2.3, 19.3.4, 19.4.3 and 19.5.3 for the different kinds of intervention models while techniques are presented in Sections 17.3 and 17.5.3 for TFN models for which there are one or more input series. After one or more intervention models are tentatively designed, MLE's can be obtained for the model parameters using the estimator described in Appendix A17.1. Automatic selection criteria such as the AIC and BIC can be employed for model discrimination purposes where the model which is ultimately selected should satisfy stringent *diagnostic checks*.

As emphasized throughout this book, all of the model construction tools should be used in an *interactive manner* by the practitioner. For instance, when deciding upon which parameters to include in an intervention model for describing a specified time series, the modeller personally examines the plotted output from a number of identification techniques. Because the output from the identification methods are usually simple to interpret, an appropriate model can usually be easily designed. Nevertheless, the practitioner must exercise a lot of *common sense* when systematically designing an intervention model with the assistance of scientific tools. The water quantity applications of Sections 19.2.4, 19.2.5, 19.3.6, 19.5.4 and 22.4.2, the temperature data application of Section 19.3.6 and also the water quality studies of Sections 19.4.5 and 22.4.2, clearly demonstrate how intervention models can be conveniently constructed by a modeller who practices both the *art and science* of model building. Finally, for a state-space representation of the intervention model, the reader can refer to Noakes (1984, Ch. 8) and Harvey (1989, Section 7.6).

Because the general intervention model is defined for the case where there is one output series, the general model in [19.5.8] is in fact a *univariate model*. This assumption is most appropriate for modelling natural time series where usually feedback is not present. For example, precipitation causes riverflows and not vice versa. Nonetheless, in some situations feedback may occur and it may therefore be necessary to use a *multivariate intervention model*. As explained by Abraham (1980) and also in Chapters 20 and 21 in this book, the multivariate model is a simple extension of the univariate case. Abraham (1980) employs a bivariate economic example to show how a multivariate intervention model can be constructed. The authors of this book would like to stress once again that practitioners should only revert to using a more complex model, such as the multivariate intervention model, when it is deemed absolutely necessary. A multivariate model is not required for any of the applications in this chapter as well as the applications in Chapters 22, 17 and 18.

Besides handling nonseasonal data, the intervention model can also be used with *seasonal data*. For the applications of Sections 19.2.5 and 19.3.6, the data are deseasonalized before intervention models are constructed. In the application in Section 19.5.4 as well as the last two applications in Section 22.4.2, covariate series in the intervention models eliminate the need for deseasonalizing the monthly series while in Section 19.4.5 deseasonalization is not required with the average monthly water quality series. When the correlation structure is dependent upon the season or group of seasons within a year, then it may be appropriate to employ the *periodic intervention model* of Section 19.6. Recall that for the periodic intervention model, a separate intervention model is fitted to each season or group of consecutive seasons for which the correlation structure is the same. Because each season possesses one output, across all the seasons the periodic intervention model can in fact be considered as a special kind of multivariate model. As noted in Section 19.6, further research is still required for developing more comprehensive model construction tools for the periodic intervention model. Perhaps a Kalman filtering approach for the periodic intervention model as well as the model in [19.5.8] may be useful. However, the periodic version of the intervention model requires many more parameters than the model in [19.5.8] and hence the practitioner should only use this model when it is deemed necessary and there are sufficient data. Simplified versions of the periodic intervention model are discussed in Sections 19.6.1 and 14.6.3, while a water quantity application is presented in Section 19.6.2.

An alternative, but related approach to studying intervention analysis, is presented by Box and Tiao (1976). Subsequent to the date of occurrence of a known intervention, a model, such as an ARIMA model, can be calibrated to the time series being considered. This calibrated model, which is appropriate for modelling the data before the intervention, can then be used to generate forecasts starting with the time when the intervention comes into effect. By comparing the forecasts with what actually occurs on and after the date of the intervention, the nature of the possible changes caused by the intervention on the time series can be studied. Box and Tiao (1976) devise a χ^2 test for ascertaining whether or not the intervention created a significant change in the mean level of the series. However, this approach differs from the intervention model in this chapter because only the series before the intervention is used to calibrate the model whereas the data from both before and after the intervention are utilized for estimating the parameters in the intervention model of [19.5.8]. In addition, Box and Tiao (1976) mention various drawbacks to their *forecasting approach to intervention analysis* and because of these negative aspects, the procedure is not considered in detail in this chapter. The authors also point out that their procedure is related to, but different from, the problem of sequential surveillance of routine forecasting schemes where one-step ahead forecast errors are available sequentially and a continuous monitoring is carried out to detect possible changes in the model. Other trend detection techniques which can be employed for discovering unknown interventions are discussed in Section 19.2.3.

Based upon a knowledge of the general type of time series model which will be eventually fitted to a given set of data, an appropriate *data collection* scheme can be devised. As explained in Section 19.7 for the case of an intervention model, by designing a suitable data collection system, full advantage can be taken of the inherent mathematical attributes of the model which will be used to analyze the data. This in turn will allow the maximum amount of information to be extracted from the data when the time series is analyzed using intervention analysis. Unfortunately, in practice, time series measurements are often not collected in an optimal manner. Sometimes, data are gathered at uneven time intervals where there may be relatively long periods of time for which no data are collected at all. This is especially true for environmental time series where, in addition to large gaps in the data, there may be multiple external interventions affecting the time series. In Part X, it is explained how *messy environmental data* can be analyzed using statistical techniques which include intervention analysis, parametric trend tests and regression analysis. Before this, however, multivariate ARMA models are presented next in Chapters 20 and 21 of Part IX.

PROBLEMS

19.1 In Section 19.1 documented applications of intervention analysis to a variety of different fields are referred to.

(a) Select one of the referenced case studies which is not described later in Chapter 19. Outline how the intervention analysis study was carried out and how intervention analysis assisted in obtaining an enhanced understanding of the problem so that informed decisions could eventually be made for alleviating the impacts of the intervention.

(b) In a field that is of direct interest to you, locate an article that describes an application of intervention analysis. Explain, in general, how the technique was applied and describe the main findings.

19.2 In Section 19.1, it is pointed out that it is usually not appropriate to apply the student t test to most intervention problems. After defining the student t test, explain in some detail the main situations in which the student t test can and cannot be applied. Base your arguments upon the theoretical properties of the test. Can intervention analysis be applied to the situations to which you stated the student t test could not?

19.3 For each of the following two dynamic responses, calculate the impulse response weights and steady state gain:

(a)
$$\frac{\omega_0 B^2 S_t^{(T)}}{1 - \delta_1 B - \delta_2 B^2}$$

where $S_t^{(T)}$ is the step indicator variable defined in [19.2.3],

(b) $$\frac{(\omega_0 - \omega_1 B)B^2 P_t^{(T)}}{1 - \delta_1 B}$$

where $P_t^{(T)}$ is the pulse indicator variable defined in [19.2.5].

19.4 Suppose that an intervention model is written as

$$y_t - \mu_y = \frac{\omega_0}{1 - \delta_1 B}\xi_t + \frac{(1 - \theta_1 B)}{(1 - \phi_1 B)}a_t$$

where ξ_t is the step response given in [19.2.3] such that

$$\xi_t = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}$$

and y_t is not transformed using a Box-Cox transformation. Derive the expression for obtaining both the change and percentage change in the mean level of the response series caused by the intervention.

19.5 For the intervention model written in the previous question, suppose that the original series, Y_t, is first transformed using natural logarithms to obtain y_t. Derive the expression for calculating the percentage change in the mean level for the original series.

19.6 Describe the change-detection statistic of MacNeill (1985) for discovering the parameter changes in a time series which occur at unknown times. By referring to other references given in Section 19.2.3 in the subsection on other trend detection techniques, explain how MacNeill's work has been expanded since 1985. Outline how MacNeill's change-detection statistic could be employed in a comprehensive intervention analysis study of a given set of environmental time series.

19.7 Explain how the technique of Bagshaw and Johnson (1977) works for detecting changes in a time series model.

19.8 Outline how the method of Fiorina and Maffezzoni (1979) is designed for detecting jumps in linear time-invariant systems and how you think it could be employed in discrete time.

19.9 In Section 19.2.3, a range of informative graphical procedures are suggested for detecting unknown interventions and investigating the stochastic impacts of either known or newly discovered interventions upon a given time series. Use appropriate exploratory data analysis techniques for studying the effects of a suspected intervention upon a nonseasonal time series which is of direct interest to you. Comment upon your findings.

19.10 Using the yea.ly time series from the previous problem or else another annual data set which has been subjected to an external intervention, follow the three stages of model construction described in Section 19.2.3 to fit an intervention model to the time series.

19.11 Execute problem 19.9 using a seasonal data set.

19.12 Carry out problem 19.11 for the case of a seasonal time series.

19.13 Using a representative TFN model, explain how the back forecasting method referred to in Sections 18.5.2 and 19.3.2, can be employed for data filling. Apply this procedure to an actual set of time series selected by you. Discuss the benefits and disadvantages of this type of record extension.

19.14 Briefly describe the approach of Coons (1957) for filling in missing data and point out the main advantages and drawbacks of the method. Compare Coons' technique for estimating missing observations to the intervention analysis method of Section 19.3.

19.15 Explain the main ideas underlying seasonal adjustment procedures to data filling, such as the one presented in Section 22.2. In what kinds of situations would you use this procedure and what are the major assets and drawbacks of the method?

19.16 Using mathematical equations when necessary, outline the approach of Brubacher and Wilson (1976) for estimating missing observations. By comparing the technique to other data filling methods, explain the advantages and drawbacks of their procedure.

19.17 By employing mathematical equations, briefly describe the EM algorithm of Dempster et al. (1977) for obtaining MLE's of the parameters of a model being fitted to an incomplete data set. Discuss the strengths and weaknesses of their procedure. Point out any commonalities between their approach and the one developed by Jones (1980) for the case of ARMA models.

19.18 Select a nonseasonal time series which is of direct interest to you and has not been impacted by external interventions. Remove six observations at different locations in the series and then employ the intervention analysis approach to data filling of Section 19.3 to estimate the missing observations. By utilizing equations, graphs and the SE's of the estimates for the missing values, comment upon the accuracy and quality of your results.

19.19 Follow the instructions of problem 19.18 for the case of a seasonal time series.

19.20 Choose a nonseasonal time series which has been impacted by one external intervention. Develop the most appropriate intervention model to fit to this data set by following the three stages of model construction explained in Section 19.2.3. Next, remove any two observations before the intervention data and one after the intervention. Employ the intervention model of Section 19.4 to simultaneously model the impact of the intervention and estimate the missing data points. Interpret and discuss your main results. Does the intervention model, for example, provide reasonable estimates for the missing observations?

19.21 Repeat the instructions of problem 19.20 for the case of a seasonal time series.

19.22 Select a set of nonseasonal time series for which you have at least one response series that has been affected by an external intervention and at least one covariate series that has not been acted upon by an intervention. The output or response series, for example, may be average annual riverflows whereas the input or covariate series may be average yearly precipitation. Follow the three stages of model construction to develop an intervention model to describe the data set. Next remove any four data points from the response series. Then fit the general intervention model from [19.5.8] to the resulting set of time series so that missing observations can be simultaneously estimated along with the effects of the intervention and covariate series upon the response. Clearly explain how you modelled the data, point out any insights that attracted your attention, and calculate the change in the mean level of the response series due to the intervention.

19.23 Repeat the instructions of problem 19.22 for the case of a set of seasonal series.

19.24 Design an intervention model that allows for the noise term to change before and after an intervention.

19.25 Write down the finite difference equations for the periodic version of the general intervention model in [19.5.8]. Discuss the advantages and drawbacks of the periodic intervention model.

19.26 Formulate the equations for a multivariate intervention model. Discuss the types of situations where this multivariate model could be applied and explain its weaknesses and strengths.

19.27 By referring to Lettenmaier et al. (1978) describe the simulation experiments that these authors carried out to arrive at their suggestions for data collection.

19.28 By employing equations when necessary, summarize Box and Tiao's forecasting approach to intervention analysis.

REFERENCES

CUMULATIVE SUM TECHNIQUE

Barnard, G. A. (1959). Control charts and stochastic processes. *Annals of Mathematical Statistics*, 16:236-253.

Lucas, J. M. (1985). Control data cusums. *Technometrics*, 27:129-144.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41:100-114.

Woodward, R. H. and Goldsmith, P. L. (1964). Cumulative sum techniques. In *Mathematical and Statistical Techniques for Industry*, Monograph No. 3, Imperial Chemical Industries Ltd. Oliver and Boyd, Edinburgh.

DATA COLLECTION

Lettenmaier, D. P. (1978). Design considerations for ambient stream water quality monitoring. *Water Resources Bulletin*, 4(4):884-902.

Lettenmaier, D. P., Hipel, K. W. and McLeod, A. I. (1978). Assessment of environmental impacts, Part two: Data collection. *Environmental Management*, 2(6):537-554.

DATA SETS

Hurst, H. E., Black, R. P. and Simaika, Y. M. (1946). The Nile Basin, Volume VII, The future conservation of the Nile. Ministry of Public Works, Physical Department Paper No. 51, S. O. P. Press, Cairo, Egypt.

ESTIMATING MISSING DATA

Anderson, R. L. (1946). Missing plot techniques. *Biometrics*, 2:21-47.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Journal of the Royal Statistical Society Supplementary*, 4:137-170.

Beauchamp, J. J., Downing, D. J. and Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin*, 25(5):961-975.

Bloomfield, P. (1970). Spectral analysis with randomly missing observations. *Journal of the Royal Statistical Society*, Series B, 32:369-380.

Brubacher, S. R. and Wilson, G. T. (1976). Interpolating time series with applications to the estimation of holiday effects on electricity demand. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 25(2):107-116.

Chin, D. A. (1988). Spatial correlation of hydrologic time series. *Journal of Water Resources Planning and Management*, American Society of Civil Engineers 114(5):578-593.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829-836.

Coons, I. (1957). The analysis of covariance as a missing plot technique. *Biometrics*, 13:387-405.

D'Astous, F. and Hipel, K. W. (1979). Analyzing environmental time series. *Journal of the Environmental Engineering Division*, ASCE, 105(EE5):979-992.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):1-38.

Grygier, J. C., Stedinger, J. R. and Yin, H-B. (1989). A generalized maintenance of variance extension procedure for extending correlated series. *Water Resources Research*, 25(3):345-349.

Hirsch, R. M., Slack, J. R. and Smith, R. A. (1982). Techniques for trend assessment for monthly water quality data. *Water Resources Research*, 18(1):107-121.

Jones, R. H. (1962). Spectral analysis with regularly missed observations. *Annals of Mathematical Statistics*, 32:455-461.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22(3):389-395.

Lettenmaier, D. P. (1980). Intervention analysis with missing data. *Water Resources Research*, 16(1):159-171.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Ljung, G. M. (1982). The likelihood function for a stationary Gaussian autoregressive-moving average process with missing observations. *Biometrika*, 61(1):265-268.

Marshall, R. J. (1980). Autocorrelation estimation of time series with randomly missing observations. *Biometrika*, 67(3):567-570.

McLeod, A. I., Hipel, K. W. and Camacho, F. (1983). Trend assessment of water quality time series. *Water Resources Bulletin*, 19(4):537-547.

Neave, H. R. (1970). Spectral analysis with initially scarce data. *Biometrics*, 57:111-122.

Parzen, E. (1963). On spectral analysis with missing observations and amplitude modulation. *Sankhya*, Series A, 25:383-392.

Preece, D. A. (1971). Iterative procedures for missing values in experiments. *Technometrics*, 13(4):743-753.

Scheinok, P. A. (1965). Spectral analysis with randomly missed observations: the binomial case. *Annals of Mathematical Statistics*, 36:971-977.

Wilkinson, G. N. (1958). Estimation of missing values for the analysis of incomplete data. *Biometrics*, 14(2):257-286.

HYDROLOGY

Moss, M. E. and Bryson, M. C. (1974). Autocorrelation structure of monthly streamflows. *Water Resources Research*, 10:737-744.

Saskatchewan Government (1974). *1974 Operation of the Saskatchewan River System*. Technical Report HYD-6-26, Environment Saskatchewan, Hydrology Branch.

Shalash, S. (1980a). The effect of the High Aswan Dam on the hydrological regime of the River Nile. In *The Influence of Man on the Hydrological Regime with Special Reference to Representative and Experimental Basins, Proceedings of the Helsinki Symposium*, (held in June, 1980), IAHS (International Association of Hydrological Sciences) - AISH Publication No. 130, pages 244-250.

Shalash, S. (1980b). The effect of the High Aswan Dam on the hydrochemical regime of the River Nile. In *The Influence of Man on the Hydrological Regime with Special Reference to Representative and Experimental Basins, Proceedings of the Helsinki Symposium*, (held in June, 1980), IAHS (International Association of Hydrological Sciences) - AISH Publication No. 130, pages 251-257.

Yevjevich, V. and Jeng, R. I. (1969). *Properties of Non-homogeneous Hydrologic Series*. Technical Report, Hydrology Paper No. 32, Colorado State University, Fort Collins, Colorado.

INTERVENTION ANALYSIS

Abraham, B. (1980). Intervention analysis and multiple time series. *Biometrika*, 67(1):73-78.

Baracos, P. C., Hipel, K. W. and McLeod, A. I. (1981). Modelling hydrologic time series from the Arctic. *Water Resources Bulletin*, 17(3):414-422.

Beauchamp, J. J., Downing, D. J., and Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin*, 25(5):961-975.

Bhattacharyya, M. N. and Layton, A. P. (1979). Effectiveness of seat belt legislation on the Queensland Road Toll - an Australian case study in intervention analysis. *Journal of the American Statistical Association*, 74(367):596-603.

Bilonick, R. A. and Nichols, D. G. (1983). Temporal variations in acid precipitation over New York State - What the 1965-1979 USGS data reveal. *Atmospheric Environment*, 17(6):1063-1072.

Box, G. E. P. (1974). Statistics and the environment. *Journal of the Washington Academy of Science*, 64(2):52-59.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70(349):70-79.

Box, G. E. P. and Tiao, G. C. (1976). Comparison of forecast and actuality. *Journal of the Royal Statistical Society*, Series C, 25(3):195-200.

Downing, D. J., Pack, D. J. and Westley, G. W. (1983). A diverting structure's effects on a river flow time series. *Management Science*, 29(2):225-236.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, United Kingdom.

Hipel, K. W. (1981). Geophysical model discrimination using the Akaike information criterion. *IEEE Transactions on Automatic Control*, AC-26(2):358-378.

Hipel, K. W., Lennox, W. C., Unny, T. E. and McLeod, A. I. (1975). Intervention analysis in water resources. *Water Resources Research*, 11(6):855-861.

Hipel, K. W., Lettenmaier, D. P. and McLeod, A. I. (1978). Assessment of environmental impacts, Part one: Intervention analysis. *Environmental Management*, 2(6):529-535.

Hipel, K. W. and McLeod, A. I. (1981). Box-Jenkins modelling in the geophysical sciences. In Craig, R. G. and Labovitz, M. L., editors, *Future Trends in Geomathematics*, pages 65-86. Pion, Great Britain.

Hipel, K. W. and McLeod, A. I. (1989). Intervention analysis in environmental engineering. *Environmental Monitoring and Assessment*, 12:185-201.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977a). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research*, 13(3):567-575.

Hipel, K. W., McLeod, A. I. and McBean, E. A. (1977b). Stochastic modelling of the effects of reservoir operation. *Journal of Hydrology*, 32:97-113.

McLeod, A. I., Hipel, K. W. and Camacho, F. (1983). Trend assessment of water quality time series. *Water Resources Bulletin*, 19(4):537-547.

McLeod, G. (1983). *Box-Jenkins in practice, Volume 1, Univariate Stochastic and Transfer Function/Intervention Analysis*. Gwilym Jenkins and Partners Ltd., Parkfield, Greaves Road, Lancaster, England.

Noakes, D. J. (1986). Quantifying changes in British Columbia dungeness crab (cancer magister) landings using intervention analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, 43(3):634-639.

Noakes, D. J. and Campbell, A. (1992). Use of geoduck clams to indicate changes in the marine environment of Ladysmith Harbour, British Columbia. *Environmetrics*, 3(1):81-97.

Shaw, D. T. and Maidment, D. R. (1987). Intervention analysis of water use restrictions, Austin, Texas. *Water Resources Bulletin*, 23(6):1037-1046.

Vandaele, W. (1983). *Applied Time Series and Box-Jenkins Models*. Academic Press, New York.

Whitfield, P. H. and Woods, P. F. (1984). Intervention analysis of water quality records. *Water Resources Bulletin*, 20(5):657-667.

Wichern, D. W. and Jones, R. H. (1977). Assessing the impact of market disturbances using intervention analysis. *Management Science*, 24:329-337.

TREND AND CHANGE DETECTION

Bagshaw, M. and Johnson, R. A. (1977). Sequential procedures for detecting parameter changes in a time-series model. *Journal of the American Statistical Association*, 72(359):593-597.

Brillinger, D. R. (1989). Consistent detection of a monotonic trend superposed on a stationary time series. *Biometrika*, 76(1):23-30.

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society*, Series B, 37:149-192.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics*, 35:999-1018.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829-836.

Feder, P. I. (1975). The log likelihood ratio in segmented regression. *Annals of Statistics*, 3:84-97.

Fiorina, M. and Maffezzoni, C. (1979). A direct approach to jump detection in linear time-invariant systems with application to power system perturbation detection. *IEEE Transactions on Automatic Control*, AC-24(3):428-434.

Gardner, L. A. (1969). On detecting changes in the mean of normal variates. *Annals of Mathematical Statistics*, 40:116-126.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, 56:495-504.

Jandhyala, V. K. and MacNeill, I. B. (1989). Residual partial sum limit process for regression models with applications to detecting parameter changes at unknown times. *Stochastic Processes and their Applications*, 33:309-323.

Jandhyala, V. K. and MacNeill, I. B. (1991). Tests for parameter changes at unknown times in linear regression models. *Journal of Statistical Planning and Inference*, 27:291-316.

Kennett, R. and Zacks, S. (1992). *Tracking Algorithms for Processes with Change Points*. Working Paper 92-218, The School of Management, State University of New York at Binghamton.

MacNeill, I. B. (1974). Tests for change of parameter at unknown time and distributions of some related functionals of Brownian motion. *Annals of Statistics*, 2:950-962.

MacNeill, I. B. (1978a). Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *Annals of Statistics*, 6:422-433.

MacNeill, I. B. (1978b). Limit processes for sequences of partial sums of regression residuals. *Annals of Probability*, 6:695-698.

MacNeill, I. B. (1980). Detection of changes in the parameters of periodic or pseudo-periodic systems when the change times are unknown. In S. Ikeda et al., Editors, *Statistical Climatology*, pages 183-195. Elsevier, Amsterdam, The Netherlands.

MacNeill, I. B. (1985). Detecting unknown interventions with application to forecasting hydrological data. *Water Resources Bulletin*, 21(4):785-796.

MacNeill, I. B., Tang, S. M. and Jandhyala, V. K. (1991). A search for the source of the Nile's change-points. *Environmetrics*, 2(3):341-375.

Noakes, D. J. (1984). *Applied Time Series Modelling and Forecasting*. Ph.D. Thesis, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41:100-115.

Page, E. S. (1955). A test for change in a parameter occurring at an unknown point. *Biometrika*, 42:523-527.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53:873-880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55:324-330.

Tang, S. M. and MacNeill, I. B. (1993). The effect of serial correlation on tests for parameter change at unknown time. *The Annals of Statistics*, 21(1):552-575.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston.

Wichern, D. W., Miller, R. B. and Hsu, D-A. (1976). Changes of variance in first-order autoregressive time series models - with an application. *Applied Statistics*, 25(3):248-256.

Zetterqvist, L. (1991). Statistical estimation and interpretation of trends in water quality time series. *Water Resources Research*, 27(7):1637-1648.

PART IX

MULTIPLE INPUT-MULTIPLE OUTPUT MODELS

As explained in Parts VII and VIII, in many natural systems, a single output or response variable is caused by one or more input or covariate variables. For example, riverflows are caused by physical variables which include precipitation and temperature. To model a system for which one or more variables cause another but not vice versa, the **TFN model** of Part VII can be employed. When one or more external interventions have modified the behaviour of the output series, the **intervention model** of Chapter 19 and Section 22.4 can be used. The intervention model is, in fact, a special type of TFN model.

When there is **feedback** in a system for which one variable causes another and vice versa, one must use a **multivariate model** to describe this situation. According to the definition used in the statistical literature, a multivariate time series model that is designed for handling feedback contains both multiple input and multiple output series. Although feedback is not as common as one way causality in hydrological systems, feedback can sometimes occur. A large lake, for instance, may affect local climatic conditions and thereby create precipitation which in turn increases the water level of the lake. In socio-economic systems, the phenomenon of feedback is very common. For example, unemployment may cause inflation which in turn increases unemployment. A detailed discussion of how to statistically detect various kinds of causality, including feedback, using the **residual** CCF (cross-correlation function) is presented in Section 16.2.2.

To formally model a system that contains feedback, a multivariate model must be used. In particular, Part IX of the book focuses upon the **general multivariate ARMA model** and simplifications thereof. Qualitatively, a general multivariate ARMA model can be written as

Multiple Outputs = Multiple Inputs + Multiple Noise

When interventions affect one or more of the series, the multivariate model can be easily extended to handle that situation. Furthermore, because the multivariate model describes how series influence one another over time, it is a **dynamic model**.

In Chapter 20, the general multivariate ARMA model is defined and other kinds of multivariate models that have been used in hydrology and environmental engineering are described and compared. Because the general multivariate ARMA model contains a large number of parameters, it is too cumbersome and overly complex for modelling most water resources systems problems. Nevertheless, the **TFN** and **contemporaneous ARMA (CARMA) models** constitute **two important subsets** of the general multivariate ARMA model that are well designed for effectively modelling water resources systems. As demonstrated by extensive applications for the TFN models in Chapters 17 and 18, and for the intervention models in Chapter 19 and Section 22.4, these models have widespread applicability in the environmental sciences.

The **CARMA** model is designed for modelling situations where two or more series affect one another at the same time or simultaneously. Because of this, the model is called the contemporaneous ARMA or simply CARMA model. Although CARMA models are not used as often as TFN models in water resources, they are still indispensible for modelling many types of problems. For example, two riverflow series measured at sites in two different rivers neither of which is upstream from the other may be related contemporaneously because the two measuring sites fall within the same general climatic region. Rather than separately model each of the two time series using an ARMA model, a CARMA model can more efficiently model both series together within a single mathematical framework. In Chapter 21, the CARMA model is defined and flexible model building procedures are presented. Both water quantity and quality applications confirm the great utility of CARMA models in water resources and environmental engineering.

Figure IX.1 depicts the hierarchical relationships among the dynamic models described in the book. Additionally, the figure contains the locations in the book where the definitions, model construction tools and applications for these dynamic models can be found.

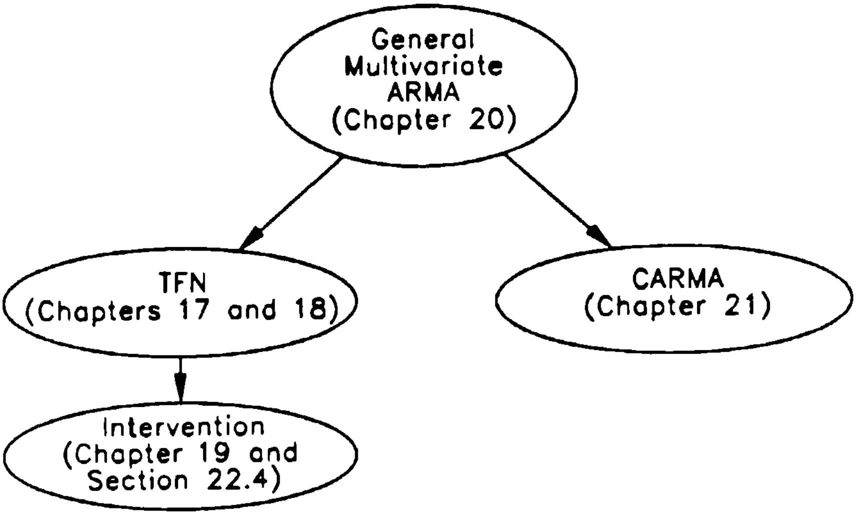

General Multivariate ARMA (Chapter 20)
TFN (Chapters 17 and 18)
CARMA (Chapter 21)
Intervention (Chapter 19 and Section 22.4)

Figure IX.1. Hiearchy of dynamic models.

CHAPTER 20

GENERAL MULTIVARIATE AUTOREGRESSIVE MOVING AVERAGE MODELS

20.1 INTRODUCTION

The term *multivariate* possesses a number of related interpretations that are used commonly by both practitioners and researchers. For example, many people consider the word multivariate to indicate that multiple variables in a system have been measured and, consequently, a multivariate model is needed to model the system. Under this definition, the TFN models of Part VII would be classified as multivariate models because the model statistically describes how one or more input variables affect the behaviour of a single output variable. Likewise, any deterministic model or mixed deterministic-stochastic model that formally describes the relationships among at least two physical variables can be thought of as being a multivariate model.

Because this book deals mainly with stochastic or time series models, the statistical definition of multivariate models is utilized. In particular, as noted in the preface to Part IX, a *general multivariate ARMA model* is a model that statistically describes how multiple outputs are influenced by multiple inputs and multiple noise terms. According to this statistical definition, the TFN models of Part VII and the related intervention models of Chapter 19 and Section 22.4 are not multivariate models. As a matter of fact, since these models possess a single output variable they are statistically classified as being univariate models.

The many time series applications presented throughout this book firmly establish the fact that the scientific community clearly recognizes the importance of time series modelling in water resources and environmental engineering. Indeed, as the demand for water continues to increase and more and more of the natural environment is altered due to industrialization and other land use changes, greater emphasis will be placed upon using more flexible systems sciences methodologies to assist decision makers in water resources (see Sections 1.2 to 1.5). To better understand how man's activities affect the environment, extensive measurements will have to be taken of a wide range of water quality variables, riverflows and lake levels, meteorological phenomena, as well as many other kinds of variables. The resulting vast amounts of data will have to be stored, processed and transferred using extensive computer networks. This in turn means that the need for having comprehensive multivariate models for describing multiple time series will continue to expand. In fact, the foregoing scenario is what Prof. V. Yevjevich considers to be the *major challenge for hydrological research* (personal communication from Prof. V. Yevjevich during the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes held at Colorado State University, Fort Collins, Colorado, July 15 to 17, 1985). Due to the current and expanding importance of multivariate analysis in hydrology, the organizers of the Fourth International Hydrology Symposium selected the theme of their conference to be *Multivariate Analysis of Hydrologic Processes*. Within this text, Chapters 20 and 21 put multivariate modelling into proper perspective and present attractive kinds of stochastic multivariate models for use in practical applications.

Stochastic or time series models are not the only type of multivariate models that can be used in water resources for modelling more than one physical variable at the same time. A *conceptual model* constitutes a *deterministic model* that is specifically designed to mathematically simulate the physical processes involved in the hydrological cycle. When studying a given problem, a scientist should employ a type of multivariate model which he or she feels is most useful and realistic. In some cases, a scientist may utilize a physically based (i.e., conceptual) model which he thinks can explain certain deterministic aspects of a natural system. After removing the portion of the data which can be explained using a physical model, the scientist can then model what is left over using a stochastic model. The overall model is referred to as a *mixed deterministic-stochastic model*. Applications of stochastic, deterministic and mixed deterministic-stochastic models to hydrological systems are given in the Proceedings of the Fourth International Hydrology Symposium (Shen et al., 1986). Within the Proceedings, a keynote paper on stochastic research in multivariate analysis is presented by Hipel (1986). In a specially edited Monograph on *Time Series Analysis in Water Resources* (Hipel, 1985b), Salas et al. (1985) review and compare alternative approaches for modelling multiple water resources time series.

Conceptual models can possess a number of common problems. In particular, they are often very complex and have a large number of parameters related to physical phenomena, all of which must be calibrated (Tong et al., 1985). Furthermore, due to the great complexity of natural systems, the conceptual models are, like other models, only rough approximations to reality. As demonstrated by a case study in Section 18.3 and also by Thompstone et al. (1985a), a simple stochastic TFN model forecasts more accurately than a cumbersome conceptual model which is very expensive to maintain and calibrate.

Even though most time series models were not originally designed to reflect the behaviour of physical phenomena, a *physical basis* to these models can often be justified. For instance, as explained by Salas and Smith (1981) and also in Section 3.6, a particular conceptual model of a watershed leads to ARMA streamflows and ARMA groundwater storage. Further discussions regarding physically based models are give by Klemes (1978). Yevjevich and Harmancioglu (1985) stress the importance of linking stochastic models with physically consistent properties of any particular water resources time series.

Many time series analysis approaches to multivariate modelling fall within the general framework of multivariate ARMA models. Consequently, in the next section the general ARMA multivariate is defined, while in Section 20.3 model construction is discussed and modelling limitations are clearly pointed out. Subsequent to this, an historical overview of the development of multivariate ARMA time series modelling in water resources is presented. Part of this historical evolution leads to the conclusion that the contemporaneous ARMA (CARMA) and TFN models constitute the two subclasses of the general family of multivariate ARMA models that are suitable for use in practical applications. Accordingly, as shown in Figure IX.1, TFN, intervention, and CARMA models are studied in depth in this book in Chapters 17 and 18, Chapter 19 and Section 22.4, and Chapter 21, respectively. Following the historical summary, of the development of multivariate ARMA models, other families of multivariate models are discussed in Section 20.5. In Section 20.5.2, the ongoing debate regarding the relative usefulness and philosophical foundations of disaggregation and aggregation models is described. Additional classes of models referred to in Section 20.5 include nonGaussian, nonlinear, fractional differencing, frequency domain, pattern recognition, and nonparametric models, in Sections 20.5.3 to 20.5.8, respectively. In the conclusions, a wide variety of challenging problems are suggested for future research projects in multivariate time series modelling in water resources and environmental engineering.

20.2 DEFINITIONS OF MULTIVARIATE ARMA MODELS

20.2.1 Introduction

After defining the general family of multivariate ARMA models, the TFN and CARMA classes of models are defined as subsets of this general family. As described in Section 20.3.2, some rather cumbersome model construction techniques are available for use with general multivariate ARMA models. However, limitations on using general multivariate ARMA models in water resources applications are clearly pointed out in Section 20.3.1. To overcome these drawbacks, subfamilies of multivariate ARMA models are suggested for use in hydrology. In particular, the CARMA model described in detail in Chapter 21 is recommended for modelling multiple time series when the series are contemporaneously correlated with one another at a given time but not at lags other than zero. To model a single response series which is driven by one or more covariate series plus a noise component, the TFN family of models of Part VII can be used. As explained in Section 20.4, it is interesting to note how the development of multivariate modelling in water resources converged over a period of two decades to the conclusion that CARMA and TFN models are the most appropriate kinds of multivariate ARMA models to use in practical hydrological applications.

20.2.2 Definitions

General Multivariate ARMA model

Let a set of k time series be represented at time t by the vector

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$$

where the vector of the theoretical means for $\mathbf{Z}_t$ is given by $\mu = (\mu_1, \mu_2, \ldots, \mu_k)^T$ and the superscript T stands for the transpose of a vector. If the AR (autoregressive) order is p and the MA (moving average) order is q, the *general k-dimensional multivariate ARMA(p,q) model* can be conveniently written as

$$(\mathbf{Z}_t - \mu) - \Phi_1(\mathbf{Z}_{t-1} - \mu) - \Phi_2(\mathbf{Z}_{t-2} - \mu) - \cdots - \Phi_p(\mathbf{Z}_{t-p} - \mu)$$
$$= \mathbf{a}_t - \Theta_1\mathbf{a}_{t-1} - \Theta_2\mathbf{a}_{t-2} - \cdots - \Theta_q\mathbf{a}_{t-q} \qquad [20.2.1]$$

where

$$\Phi_i = \begin{bmatrix} \phi_{11i} & \phi_{12i} & \ldots & \phi_{1ki} \\ \phi_{21i} & \phi_{22i} & \cdots & \phi_{2ki} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \phi_{k1i} & \phi_{k2i} & \cdots & \phi_{kki} \end{bmatrix}$$

is the ith AR parameter matrix of order $k \times k$ for $i = 1, 2, \ldots, p$;

$$\Theta_i = \begin{bmatrix} \theta_{11i} & \theta_{12i} & \dots & \theta_{1ki} \\ \theta_{21i} & \theta_{22i} & \cdots & \theta_{2ki} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \theta_{k1i} & \theta_{k2i} & \cdots & \theta_{kki} \end{bmatrix}$$

is the ith MA parameter matrix of order $k \times k$ for $i = 1,2,\ldots,q$; $\mathbf{a}_t = (a_{t1}, a_{t2}, \ldots, a_{tk})^T$, is the k-dimensional vector of innovations for $\mathbf{Z}_t$ at time t. Because the vectors have the form $(\mathbf{Z}_t - \mu)$ in [20.2.1], the multivariate ARMA model is often referred to as a vector ARMA model. When man-induced or natural interventions affect one or more of the multiple series in [20.2.1], the model can be easily extended to handle multiple interventions (Abraham, 1980).

A more compact format for writing the model in [20.2.1] is given by

$$\Phi(B)(\mathbf{Z}_t - \mu) = \Theta(B)\mathbf{a}_t \qquad [20.2.2]$$

where B is the backward shift operator defined by $B^k\mathbf{Z}_t = \mathbf{Z}_{t-h}$, $\Phi(B) = \mathbf{I} - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$ is the AR operator of order p where $\mathbf{I}$ is the identity matrix of order $k \times k$, and $\Theta(B) = \mathbf{I} - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_q B^q$ is the MA operator of order q.

There are a number of assumptions underlying the linear multivariate ARMA(p,q) model given in [20.2.1] or [20.2.2]. To start with, it is assumed that the innovations given by $\mathbf{a}_t$ are identically independently distributed (IID) vector random variables with a mean of zero and variance covariance matrix Δ. In order to obtain MLE's (maximum likelihood estimates) of the parameters and also design sensitive diagnostic checks, for practical applications it is necessary to invoke the normality assumption so that the innovations are normally independently distributed (NID) and hence $\mathbf{a}_t \approx \text{NID}(\mathbf{0},\Delta)$. Finally, to permit the model in [20.2.1] or [20.2.2] to be stationary and invertible, the zeroes of the determinant equations $|\Phi(B)| = 0$ and $|\Theta(B)| = 0$, respectively, must lie outside the unit complex circle.

Example: Consider a multivariate ARMA(1,1) model possessing two variables contained in the vector

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2})^T$$

having theoretical means given by $\mu = (\mu_1, \mu_2)^T$. Because there are two variables, the multivariate model used to describe mathematically the relationship between the two variables is called a bivariate model. Following [20.2.1], the bivariate ARMA(1,1) model is written as

$$(\mathbf{Z}_t - \mu) - \Phi_1(\mathbf{Z}_{t-1} - \mu) = \mathbf{a}_t - \Theta_1 \mathbf{a}_{t-1}$$

where

$$\Phi_1 = \begin{bmatrix} \phi_{111} & \phi_{121} \\ \phi_{211} & \phi_{221} \end{bmatrix}$$

is the AR parameter matrix;

$$\Theta_1 = \begin{bmatrix} \theta_{111} & \theta_{121} \\ \theta_{211} & \theta_{221} \end{bmatrix}$$

is the MA parameter matrix; $\mathbf{a}_t = (a_{t1}, a_{t2})$ is vector of innovations containing IID random variables. Substituting the AR and MA matrices into the bivariate ARMA model produces

$$\begin{bmatrix} Z_{t1} - \mu_1 \\ Z_{t2} - \mu_2 \end{bmatrix} - \begin{bmatrix} \phi_{111} & \phi_{121} \\ \phi_{211} & \phi_{221} \end{bmatrix} \begin{bmatrix} Z_{t-1,1} - \mu_1 \\ Z_{t-1,2} - \mu_2 \end{bmatrix} = \begin{bmatrix} a_{t1} \\ a_{t2} \end{bmatrix} - \begin{bmatrix} \theta_{111} & \theta_{121} \\ \theta_{211} & \theta_{221} \end{bmatrix} \begin{bmatrix} a_{t-1,1} \\ a_{t-1,2} \end{bmatrix}$$

After matrix multiplication, the two component equations of the bivariate model are

$$Z_{t1} - \mu_1 - \phi_{111}(Z_{t-1,1} - \mu_1) - \phi_{121}(Z_{t-1,2} - \mu_2) = a_{t1} - \theta_{111} a_{t-1,1} - \theta_{121} a_{t-1,2}$$

$$Z_{t2} - \mu_2 - \phi_{211}(Z_{t-1,1} - \mu_1) - \phi_{221}(Z_{t-1,2} - \mu_2) = a_{t2} - \theta_{211} a_{t-1,1} - \theta_{221} a_{t-1,2}$$

TFN Model

Because of the great importance of TFN models in water resources, these models are studied in depth in Chapters 17 and 18 while the closely related intervention models are entertained in Chapter 19 and Section 22.4. As noted earlier, the TFN model is a subset of the multivariate ARMA model in [20.2.1]. In particular, when the AR and MA parameter matrices in [20.2.1] are either all upper or else lower triangular, the model is called a TFN model. For the case where the matrices are lower triangular, the *TFN model* is defined following [20.2.1] as

$$(\mathbf{Z}_t - \mu) - \Phi_1(\mathbf{Z}_{t-1} - \mu) - \Phi_2(\mathbf{Z}_{t-2} - \mu) - \cdots - \Phi_p(\mathbf{Z}_{t-p} - \mu)$$
$$= \mathbf{a}_t - \Theta_1 \mathbf{a}_{t-1} - \Theta_2 \mathbf{a}_{t-2} - \cdots - \Theta_q \mathbf{a}_{t-q} \qquad [20.2.3]$$

where the ith AR parameter matrix is

$$\Phi_i = \begin{bmatrix} \phi_{11i} & 0 & 0 & \cdots & 0 \\ \phi_{21i} & \phi_{22i} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \phi_{k1i} & \phi_{k2i} & \phi_{k3i} & \cdots & \phi_{kki} \end{bmatrix}$$

and the ith MA parameter matrix is written as

$$\Theta_i = \begin{bmatrix} \theta_{11i} & 0 & 0 & \cdots & 0 \\ \theta_{21i} & \theta_{22i} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \theta_{k1i} & \theta_{k2i} & \theta_{k3i} & \cdots & \theta_{kki} \end{bmatrix}$$

Salas et al. (1985) refer to the TFN model in [20.2.3] as a triangular relationship because Z_{t1} only depends on its own past, Z_{t2} depends on its own past plus the present and past of Z_{t1}, Z_{t3} is dependent on its own past and on the present and past of Z_{t1} and Z_{t2}, and so on. In the final component equation in [20.2.3], Z_{tk} depends on its own past and on the present and past of $Z_{t1}, Z_{t2}, \ldots, Z_{tk-1}$. This means that the single output Z_{tk} depends upon the input variables

$Z_{t1},Z_{t2},\ldots,Z_{tk-1}$. Recall from Section 17.5.2 that this definition constitutes, in fact, a TFN model. By appropriate algebraic manipulations, one can easily demonstrate that the TFN model in [20.2.3] is equivalent to the more convenient form for writing the TFN model given in [17.5.3]. Comprehensive model building procedures for constructing TFN models and numerous water resources applications are presented in Chapters 17 and 18. Because a single output variable is dependent upon multiple input variables, a TFN model can be statistically classified as a univariate model.

Camacho et al. (1986) provide a simple example to demonstrate when a TFN model would clearly be selected over a general multivariate ARMA model. More specifically, when dealing with unregulated multisite hydrological systems, it can be argued that the general multivariate ARMA model would never be required to model the data and that a TFN model with only upper or lower triangular parameters will always be appropriate. To illustrate this fact, consider for simplicity the three-station riverflow system shown in Figure 20.2.1. It is clear from the nature of the system that only flows located upstream of any given station will influence the flows at that station. Therefore, if the vector of flows at time t is $\mathbf{Z}_t = (Z_{t1},Z_{t2},Z_{t3})^T$, the parameter matrices of the model in [20.2.1] or [20.2.2] will contain only (possible) nonzero elements at entries (1,1), (2,2), (3,1), (3,2) and (3,3). No other entry in the matrix should be allowed to be different from zero. For example, if the (1,3) element of a matrix were permitted to be nonzero, it would imply that flows at Station 1, Z_{t1} would be written as a linear combination of past values of Z_{t1}, past values of Z_{t3} and some error terms. This, of course, would not have any physical meaning. It is easy to see that the resulting matrices of the model are lower triangular. The same argument can be extended to more complex systems.

The simple example presented above shows that when the physical restrictions of the system are taken into consideration in the formulation of the model, it is possible to substantially reduce the number of parameters. The benefits of such a reduction can be appreciated by looking at the precision of the parameters estimates. Suppose, for example, that the bivariate series $\mathbf{Z}_t = (Z_{t1},Z_{t2})$ is modelled as a general multivariate ARMA(1,0) when in fact a CARMA (1,0) would suffice. It is shown by Camacho et al. (1985a) and also in Section 21.5 that the variances of the estimated parameters obtained using the CARMA model are always smaller than the ones obtained using the full multivariate model and that such reductions may be well over 50%.

CARMA Model

A *CARMA model* is obtained from [20.2.1] or [20.2.2] when all of the parameter matrices are diagonal. Consequently, for the AR and MA matrices given by Φ_i and Θ_i, respectively, the elements $\phi_{jmi} = 0$ and $\theta_{jmi} = 0$ for $j \neq m$ and $i = 1,2,\ldots,k$. In Chapter 21, the CARMA (p,q) model is written out in full in [21.2.1] as well as [21.2.4]. Because this parsimonious model implies a contemporaneous relationship among the concurrent multivariate observations or, equivalently, the multivariate innovations which occur at the same time t, it is referred to as a contemporaneous ARMA model. Furthermore, since a CARMA model contains multiple output series, it constitutes a multivariate model. Chapter 21 of this book is entirely devoted to this useful and interesting class of models. A comprehensive set of model construction techniques given in Section 21.3 allows CARMA models to be conveniently applied to practical problems and the water resources applications in Section 21.5 demonstrate the utility of these models.

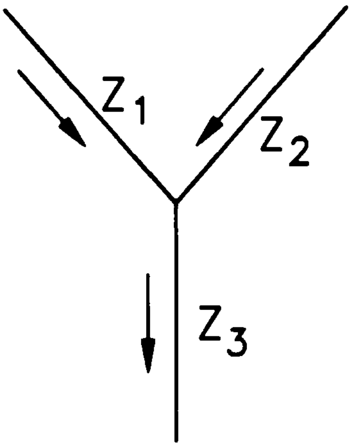


Figure 20.2.1. A three-station riverflow system where Z_i represents measurements at station i.

20.3 CONSTRUCTING GENERAL MULTIVARIATE ARMA MODELS

20.3.1 Limitations

As noted by authors such as Salas et al. (1985), Camacho et al. (1985a, 1986, 1987a,b,c) and Hipel (1986), there are two major drawbacks for using the general multivariate ARMA in [20.2.1] and [20.2.2] for applications in water resources. First of all, because the number of parameters increases exponentially with the dimensionality of the model, the multivariate ARMA model is very complicated and possesses too many parameters. Secondly, a comprehensive set of operational and simple model building techniques are not available for constructing multivariate ARMA models by following the identification, estimation and diagnostic check stages of model construction. As a result, one cannot assume the most general form of the multivariate ARMA model to begin with and employ construction techniques to identify an appropriate model to parsimoniously describe the data set under consideration.

To overcome the foregoing problems, CARMA and TFN models can be employed. Both of these subfamilies of models contain far fewer parameters than the cumbersome general multivariate ARMA model. Additionally, a wide range of flexible model building techniques are now readily available for use with each of these two classes of models. As described in detail in Chapters 17 and 19, comprehensive model building techniques are available for conveniently constructing TFN models and the closely related intervention models, respectively. In Chapter 21, flexible model construction methods are presented for building CARMA models.

As argued in this book as well as by authors such as Salas et al. (1980), Salas et al. (1985), Camacho et al. (1985, 1986, 1987a,b,c) and Hipel (1986), the physical properties of hydrological systems often dictate that TFN and CARMA models are the proper types of multivariate models to use in practice. Nonetheless, in some water resources applications which may, for example, require the use of socio-economic data, it may be necessary to employ a multivariate ARMA model that does not fall within the TFN or CARMA categories. Consequently, the purpose of this section and Appendix A20.1 is to outline some of the model construction methods that can be used for building general multivariate ARMA models. However, the reader should bear in mind that due to the complexity of the vector ARMA model, the model building methods are unwieldy and are not as flexible or comprehensive as those currently available for building TFN or CARMA models.

20.3.2 Model Construction

Introduction

As is the case for all of the families of models entertained in this book, constructing multivariate ARMA models is effected by adhering to the three iterative stages consisting of identification, estimation and diagnostic checking. Because the concept of causality described in Section 16.2 is closely linked to the identification of not only TFN models (Section 17.3.1) and CARMA models (Section 21.3), but also general multivariate ARMA models, this idea is briefly discussed next. Following the summary given by Camacho et al. (1986), a review of recent model building procedures is then given. When modelling a set of seasonal time series using a multivariate ARMA model, one of the approaches outlined in Section 20.3.3 may be used.

Causality

As explained in Section 16.2.1, Granger (1969) defined causality between two time series in terms of predictability. In particular, a variable X causes another variable, Y, with respect to a given universe or information set that included X and Y, if the present Y can be better predicted by using past values of X than by not doing so, all other relevant information (including the past of Y) being used in either case. Causality from Y to X can be defined in the same way. Feedback occurs when X causes Y and Y also causes X.

To determine the type of causality relationship that exists between X and Y, the properties of the residual CCF are examined. Following the more detailed explanation given in Section 16.2.2, let the sequences of the observations for two variables be represented by the time series X_t and Y_t, respectively. These series can be prewhitened by fitting ARMA models to the series and obtaining the white noise residuals u_t and v_t in [16.2.3] and [16.2.4] for X_t and Y_t, respectively. When required for rectifying problems with non-normality and/or heteroscedasticity in the ARMA model residuals, X_t or Y_t can be transfromed using the Box-Cox transformation in [3.4.30] prior to prewhitening. Subsequent to prewhitening, the residual CCF, written as $\rho_{uv}(k)$ at lag k between u_t and v_t can be considered by using [16.2.5]. In addition to reflecting the type of linear dependence between u and v and consequently between X and Y, $\rho_{uv}(k)$ gives the kind of causality relationship between these variables for linear systems. As explained by Pierce and Haugh (1977) and summarized in Table 16.2.1, there are many possible types of causal interactions between X and Y which can be characterized by the properties of $\rho_{uv}(k)$. If X and Y are independent, $\rho_{uv}(k) = 0$ for all k and, hence, it would not be appropriate to develop any kind of multivariate ARMA model to link these two variables. When there is unidirectional causality such that X causes Y and Y does not cause X, it can be proven that $\rho_{uv}(k) \neq 0$ for some $k > 0$, and $\rho_{uv}(k) = 0$ for either all $k < 0$ or else all $k \leq 0$. For this situation, the most appropriate kind of multivariate ARMA model to link the input or covariate series X_t with the output or response Y_t is a TFN model defined in [17.2.5]. If X and Y are only related instantaneously, $\rho_{uv}(0) \neq 0$ and $\rho_{uv}(k) = 0$ for all $k \neq 0$. When this is the case, a CARMA model in [21.2.1] or [21.2.4] can be used to mathematically describe the contemporaneous linear dependence between X_t and Y_t. Finally, if there is feedback and hence X causes Y and vice versa, $\rho_{uv}(k) \neq 0$ for some $k > 0$ and for some $k < 0$. Feedback between two or more variables can be modelled using the multivariate ARMA model in [20.2.1] or [20.2.2].

In practical applications, one examines the sample residual CCF in [16.2.6] for X_t and Y_t to identify the type of linear dependence between X and Y. Besides detecting causality between two series, the sample residual CCF can be used for identifying time series models to mathematically describe the dynamic linkage between X_t and Y_t. In addition to output from other identification techniques, this information can then be used for identifying the appropriate kind of multivariate model to fit to the series. The manner in which the sample residual CCF can be used to design TFN and CARMA models is explained in Sections 17.3.1 and 21.3, respectively. In the next three subsections, model construction methods are presented for building general multivariate ARMA models.

Identification

As shown by applications in Chapter 5, selecting the order of a simple univariate ARMA model can sometimes be challenging. For the multivariate ARMA case, model identification is far more difficult (Tiao and Tsay, 1983a; Tjostheim and Paulsen, 1982; Jenkins and Alavi, 1981; Tiao and Box, 1981). Different identification procedures have been advocated in the literature. For example, Tiao and Box (1981) and Jenkins and Alavi (1981) have extended the use of the sample CCF and the PACF to identify the order of the AR and MA operators of a multivariate process. Tiao and Tsay (1983a,b) have proposed the use of the extended sample cross correlation function (ESCCF) to identify the order of general multivariate ARMA(p,q) models. Newbold and Hotopp (1984) have expanded a two-step procedure given by Hannan and Rissanen (1982) to the multivariate case in order to identify and model. These procedures are based on the calculation of simple sample statistics and stepwise regressions. Another technique based on the estimation of heavily parameterized ARMA models and the use of a consistent information criterion (Quinn, 1980; Hannan and Quinn, 1979) has also been proposed. The difficulty of applying such procedures clearly lies in the high computational costs.

In Appendix A20.1, the following three identification methods are described:

1. sample CCF matrix,
2. sample PACF matrix,
3. ESCCF matrix.

Because the theoretical CCF matrix cuts off for pure multivariate MA processes, the sample CCF matrix can be used to identify when a multivariate MA model is required (Tiao and Box, 1981; Jenkins and Alavi, 1981). This is similar to the manner in which the sample ACF in [2.5.9] is employed for detecting when a pure MA univariate model is needed (see Section 5.3.4). The theoretical PACF matrix truncates for a pure multivariate AR process and therefore the sample PACF matrix can be used for finding out when a multivariate AR model is required to model a given data set (Tiao and Box, 1981; Jenkins and Alavi, 1981). This is similar to the way in which the sample PACF in Section 3.2.2 can be used for designing a pure univariate AR model (see Section 5.3.5). When both MA and AR parameters are needed in a multivariate ARMA model, the ESCCF matrix can be utilized for ascertaining the orders of the MA and AR parameter matrices (Tiao and Tsay, 1983a,b; Tsay and Tiao, 1984).

Estimation

The likelihood function of the general multivariate ARMA(p,q) model has been given by Nicholls and Hall (1979), Hillmer and Tiao (1979), and Wilson (1973). Conditional and exact likelihood estimators have been proposed, by these authors. It has been shown by Hillmer and Tiao (1979) that if the determinant of the moving average operator has one or more zeroes close to the unit circle, the exact likelihood should be employed. Algorithms to evaluate the likelihood function have been proposed by Hall and Nicholls (1980) and by Ansley and Kohn (1983), who discussed the use of the Kalman filter to incorporate the case of missing or aggregated data. Shea (1989) provided a computer program for calculating the exact likelihood of a multivariate ARMA(p,q) model. For estimating the parameters of TFN and CARMA models, algorithms which are more computationally efficient can be employed. When obtaining MLE's for the parameters of TFN and intervention models, the estimator given in appendix A17.1 can be utilized. For CARMA models, Camacho (1984) and Camacho et al. (1985a, 1987a,b) have developed a computationally efficient algorithm to estimate the parameters of the model. They have also extended the algorithm to include the case of CARMA models with unequal sample sizes. Their algorithm is described in Section 21.3.3.

Diagnostic Checking

As pointed out in Section 20.2.2 with equations [20.2.1] and [20.2.2], the innovations of the general multivariate ARMA model are assumed to be NID. Diagnostic checks should be executed to insure that the key residual assumptions are satisfied. In particular, to determine if the residuals are white one can employ the sample CCF and sample PACF described in Appendix A20.1. Alternatively, one can use the modified Portmanteau test of Li and McLeod (1981) for whiteness checks. To verify that the normality assumption is satisfied, one can utilize the multivariate normality tests proposed by Royston (1983). Finally, one could develop multivariate extensions of the constant variance tests of Section 7.5 to make sure that the residuals are not heteroscedastic.

If the residuals are not white, the multivariate model must be appropriately redesigned. When the residuals are not approximately normally distributed and/or homoscedastic, one may wish to transform one or more of the variables using an appropriate transformation such as the Box-Cox transformation in [3.4.30]. Following this, the parameters of the multivariate model can be estimated for the transformed series.

20.3.3 Seasonality

The general multivariate ARMA model given in [20.2.1] and [20.2.2] is defined for handling nonseasonal time series. Likewise, the model construction techniques of Section 20.3.2 are explained for the nonseasonal case. When a set of seasonal time series are to be modelled using a multivariate model, one of the two useful approaches described below can be used.

Deseasonalized Multivariate Model

A commonly used procedure for modelling seasonal data is to first deseasonalize each time series using the *deseasonalization* methods defined in [13.2.2] or [13.2.3]. As explained in Section 13.3.3, to decrease the number of parameters required for deseasonalization a Fourier series approach can be utilized. Subsequent to deseasonalization, the most appropriate type of nonseasonal multivariate model in [20.2.1] can be fit to the set of deseasonalized series by following the model construction procedures presented in Section 20.3.2 as well as Appendix A20.1.

Periodic Multivariate Model

An assumption underlying the deseasonalized model is that the correlation structure among seasons is the same throughout the year. To allow for a seasonally varying correlation structure, *periodic models* can be employed. As described in depth in Chapter 14, two popular periodic models are the PAR (periodic autoregressive) and PARMA (periodic ARMA) models. When fitting a PAR model to a single seasonal series, a separate AR model is designed for each season of the year. In a similar manner, a PARMA model consists of having a separate ARMA model for each season of the year. Within hydrology, PAR modelling dates back to the research of Thomas and Fiering (1962) who proposed a specialized type of PAR model whereby the order of the AR operator for each season is fixed at unity. More recently, authors such as Salas et al. (1980) and Thompstone et al. (1985a,b) have suggested that the order of the AR operator for each season be properly identified. Model construction techniques that can be employed with PAR models are presented in Sections 14.3 and 14.5.3 while PARMA modelling methods are discussed in Section 14.7. Because there is a separate model for each season of the year, periodic models can be considered to be special types of multivariate models (Salas et al., 1985; Vecchia et al., 1983; Vecchia, 1985a,b).

PAR and PARMA models can be considered as the periodic extensions of nonseasonal AR and ARMA models, respectively, for modelling seasonal data. Similar to the PAR and PARMA models described in Chapter 14, a periodic multivariate model would essentially consist of having a separate multivariate model for each season of the year. Salas and Pegram (1978) define the periodic version of multivariate ARMA(p,0) models while, Salas et al. (1980) present the periodic extension of the general multivariate ARMA(p,q) models given in [20.2.1] and [20.2.2]. Bartolini and Salas (1986), Haltiner and Salas (1988), Bartolini et al. (1988) and Ula (1990) investigate the statistical properties of multivariate PARMA(1,1) processes. Additionally, Salas and Abdelmohsen (1993) devise an initialization procedure for generating univariate and multivariate PAR(1), PAR(2) and PARMA(1,1) processes. Their approach is, in fact, the PARMA version of the WASIM2 simulation algorithm presented in Section 9.4 for simulating with ARMA models. In Section 19.6.3 and also in the paper by Hipel and McLeod (1981), specific kinds of periodic TFN and intervention models are proposed.

Because a periodic multivariate ARMA model possesses many more parameters than the complex nonperiodic version, one must devise ways to decrease the number of model parameters. Following the approach of Thompstone et al. (1985a) described in Section 14.5, one method to reduce the size of a periodic model is to divide the year into groups of seasons where consecutive seasons having similar correlation structures are put into the same group. The periodic multivariate ARMA model used to fit to the grouped data would only have parameters that preserve the correlation relationships among the groups of data rather than the original seasons. Secondly, as proposed by Salas et al. (1980) for PAR and PARMA models, a Fourier series approach could be utilized to reduce the number of parameters required in a periodic multivariate model. A final approach to economize on the number of parameters is to adopt the procedure suggested for periodic intervention models in Section 19.6.3. Depending upon the statistical characteristics of the multiple time series being modelled, only specified components of the multivariate ARMA model would be permitted to have a periodic structure. In fact, this is what is done for the intervention models developed in Sections 19.2.5 and 22.4.2 for modelling seasonal riverflows, as well as in Section 22.4.2 for describing water quality times series that have been impacted by external interventions.

20.4 HISTORICAL DEVELOPMENT

Recently, Salas et al. (1985) presented a comprehensive review of various approaches to multivariate modelling in hydrology. A significant portion of their paper deals with research closely related to multivariate ARMA modelling of hydrological time series. Camacho et al. (1986) also put research on multivariate ARMA modelling from the hydrological and statistical literature into proper perspective. The section follows closely the historical survey given by Hipel (1986), which was presented at the Fourth International Hydrology Symposium held at Colorado State University from July 15 to 17, 1985 (Shen et al., 1986).

Research in multivariate modelling in water resources goes back to the early 1960's when researchers such as Maas et al. (1962) introduced systems sciences techniques into the field of water resources. Much of this research dealt with proposing fairly simple multivariate models, most of which are either subsets of or else closely related to the multivariate ARMA model in [20.2.1] and [20.2.2]. In the earlier research, often the exact form of the model used for fitting to a data set was specified prior to model construction. For instance, some researchers suggested using a multivariate AR(1) model while others proposed employing a multivariate ARMA(1,1) model. This type of procedure may result in using a model that does not fit the data well. Because of this, Finzi et al. (1975) found that synthetic data generated by prespecified models were inadequate in several applications. Another disadvantage of this approach is that it may cause inefficient estimation of the model parameters (Camacho et al., 1985).

In a pioneering paper, Fiering (1964) proposed a two station multivariate model to link two series, X_t and Y_t. The model of Fiering was later modified by Kahan (1974) and Lawrance (1976).

Matalas (1967) suggested a multisite AR(1) model for use in hydrology. His model preserves both the lag zero and lag one cross covariance matrices. He also pointed out that the model could be simplified by having a diagonal AR matrix and he described a parameter estimation procedure based on the method of moments. Kuczera (1987) explained how to obtain MLE's for the parameters of a multivariate AR(1) model using the EM algorithm of Dempster et al. (1977) when there are missing observations. To take into account seasonality in a time series, Young and Pisano (1968) suggested first deseasonalizing the multivariate series before fitting the Matalas model. Furthermore, they designed improved estimation procedures and suggested transformations for removing skewness in the data.

For modelling monthly multivariate data, Bernier (1971) considered a monthly multivariate AR(1) model. His model is actually a combination of the Fiering and Matalas model.

For a multivariate AR(p) model, Pegram and James (1972) proposed a moment estimation procedure to estimate the parameters and when using the model for streamflow generation they gave reasons for diagonalizing the AR matrices. A general multivariate AR(p) model with seasonally varying parameters was designed by Salas and Pegram (1978) who suggested both the methods of moments and maximum likelihood to estimate the parameters. When their model has diagonal AR matrices, it forms a periodic contemporaneous AR(p) model.

As explained in Section 10.4, the FGN (fractional Gaussian noise) model defined in [10.4.2] was developed to model long term persistence and thereby provide an explanation for the Hurst phenomenon described in Section 10.3.1. To model the multivariate version of long term persistence, Matalas and Wallis (1971) considered the multivariate fractional Gaussian noise (FGN) model for which each of the series is modelled by a univariate FGN model with contemporaneously correlated innovations. O'Connell (1974) proposed a vector ARMA(1,1) model to describe long term persistence. Canfield and Tseng (1979) studied the same model with diagonal AR and MA matrices while Lettenmaier (1980) suggested improved estimation procedures for the vector ARMA(1,1) model.

Franchini et al. (1986) developed a type of multivariate AR model which has the ability to preserve long term persistence and to reproduce the statistical properties of the seasonal flows at more than one station situated in a given river basin. They pointed out that their model is capable of maintaining the time-space correlations at the seasonal level as well as the properties of the flow volumes at the annual level.

In 1974, Mejia et al. considered the situation where the generation of synthetic hydrological sequences are obtained from a mixture of distributions. To reproduce the historical moments in the simulated data, they proposed a transformation of the moments of the historical data to be used in the estimation of the parameters of the model. This procedure appears to have little statistical justification according to Stedinger (1981) who shows that direct estimation of the moments of the transformated historical data can result in significantly better estimates of the true cross correlations.

Kottegoda and Yevjevich (1977) compared the preservation of the correlation in the generated samples of four kinds of existing two station models. Because the models produced essentially equivalent results, they concluded that one has "to apply the simplest model with the best physical justification".

Stedinger (1981) compared different approaches for estimation of the correlations in multivariate streamflow models. He concluded that "there appears to be little statistical justification to the idea that one should select a streamflow model's parameters so as to reproduce exactly the observed correlations of the flows themselves ... and perhaps the most important lesson to be learned ... is that estimates of many streamflow model parameters are inaccurate". Therefore, it "is very reasonable to use statistically efficient parameter estimation which may not exactly reproduce the observed means, variances and correlations of the historical flows" (Stedinger and Taylor (1982a).

By generalizing the methodology of Vicens et al. (1975), Valdes et al. (1977) developed a Bayesian procedure to generate synthetic streamflows for multivariate AR models. An advantage of this procedure is that it takes into account parameter uncertainty. However, Davis (1977) and McLeod and Hipel (1978c) mention drawbacks to the simulation approach of Vicens et al. (1975) for handling parameter uncertainty. As explained in Section 9.7, the method of McLeod and Hipel (1978) for simulating, using univariate ARMA models, correctly takes into account parameter uncertainty and could easily be extended to the multivariate case. As pointed out Stedinger and Taylor (1982b), to include uncertainty in the parameters of the model is very important for obtaining realistic and honest estimates of system reliability.

In 1978, Ledolter proposed that the general class of multivariate ARMA models be used in hydrology. Salas et al. (1980) suggested the CARMA models with constant or periodic parameters constitute parsimonious models that reflect the physical reality of hydrological systems. They also proposed procedures for use in model construction. Further advances in identification, estimation and diagnostic checking of CARMA models were given by Camacho et al. (1985, 1986, 1987a,b,c). Other research related to contemporaneous modelling is given by authors including Hannan (1970), Wilson (1973), Granger and Newbold (1977), Wallis (1977), Chan and Wallis (1978), Hillmer and Tiao (1979), Nicholls and Hall (1979), Risager (1980, 1981), Tiao and Box (1981), and Jenkins and Alavi (1981).

Along with model construction procedures, Cooper and Wood (1982a,b) propose the multiple input-output model. This class of models is actually equivalent to the multivariate ARMA family in [20.2.1] and [20.2.2]. The mathematical and statistical properties of the multivariate models considered by Cooper and Wood (1982a,b) were studied by Hannan and Kavalieres (1984).

As was the case for the CARMA class of models, the space-time ARMA or STARMA family of models was designed to overcome the problem of too many parameters in multivariate ARMA models (Deutsch and Ramos, 1984, 1986; Pfeifer and Deutsch, 1980; Deutsch and Pfeifer, 1981). However, Camacho et al. (1986) argued that the parameter restrictions incorporated into the STARMA model may be too severe and thereby limit the applicability of the model. Nonetheless, Adamowski et al. (1986) found the STARMA model useful for modelling eleven raingage sites located in a watershed in Southern Ontario, Canada.

Kelman et al. (1986) devised a multivariate version of a model proposed by Kelman (1980) for separately modelling the rising and falling limbs of daily hydrographs. The multivariate extension of the model follows the approach suggested by Matalas (1967).

Srikanthan (1986) proposed a multivariate model for simulating daily climatic data. Daily rainfall was simulated using a multistate first order Markov model and the remaining climatic variables were simulated using a multistate type model (Matalas, 1967; Richardson, 1981). Nasseri (1986) utilized a multivariate AR model of order one to generate hourly rainfall for a network of raingages.

Venugopal et al. (1986) used a multisite model for simulating flows of the Narmada River system in India. In particular, the HEC-4 (Feldman, 1981) and disaggregation models were employed for the synthetic generation of riverflows.

As pointed out in Section 20.2.2 the TFN and intervention group of models is actually a subset of the general multivariate ARMA family of models in [20.2.1] and [20.2.2] when the AR and MA parameter matrices are either all upper or lower triangular. The use of TFN modelling in water resources dates back to the time of Fiering (1964) who proposed a bivariate TFN model which was later modified by Lawrance (1976). In fact, because of the great importance of TFN modelling and intervention analysis in water resources and environmental engineering, Parts VII and VIII of this book deal exclusively with TFN and intervention modelling, respectively.

References regarding the theory and practice of TFN modelling are listed at the ends of Chapters 16 to 18 while references for intervention modelling are given in the final parts of Chapter 19 and 22. At the Fourth International Hydrology Sympoisum on Multivariate Analysis of Hydrologic Processes held at Colorado State University from July 15 to 17, 1985, a number of research papers were concerned with TFN modelling (Shen et al., 1986). In particular, Nicklin (1986) employed a TFN model to mathematically formulate the dependence between nonstationary irrigation diversion and return flows. In order to identify an appropriate TFN model, he suggested novel identification procedures designed to use with his particular type of problem. Delleur (1986) developed a model consisting of a mixed model for forecasting real-time flash floods. The model consisted of a nonlinear conceptual submodel for transforming the observed rainfalls into effective precipitation followed by a TFN model relating the effective rainfall to the observed flood.

20.5 OTHER FAMILIES OF MULTIVARIATE MODELS

20.5.1 Introduction

The previous section on the historical development of statistical multivariate models in hydrology dealt mainly with models closely related to the general multivariate ARMA family of models in [20.2.1] and [20.2.2]. Other classes of statistical models have also been used in hydrology. For example, Fiering (1964) introduced multivariate analysis for generating multisite streamflows using *principal component analysis*. Numerous authors, have developed and employed *regression analysis* models for use in applications such as environmental impact assessment, data filling, and synthetic streamflow generation. In fact, within Chapter 24 of this book, ways in which regression analysis can be employed for both exploratory and confirmatory data analysis purposes are explained and illustrated. In Section 24.3, a trend analysis methodology, which uses techniques such as regression analysis and nonparametric tests (Chapter 23), is presented for detecting and modelling trends in water quality time series measured in rivers.

In a perceptive paper, Yevjevich and Harmancioglu (1985) discuss the past and future of time series analysis in water resources. Some of the challenging research projects that these authors feel should be actively pursued include the proper treatment of nonGaussian, nonlinear and multivariate time series. Besides defining new univariate and multivariate models for handling the foregoing and other problems, any new models must be made fully operational by developing appropriate model construction techniques. In a sequence of nineteen invited papers written by statisticians, hydrologists and other scientists, these as well as many other challenges are met head on by many original research contributions (Hipel, 1985b). As pointed out in Section 1.1, because of the great importance of time series analysis as well as other statistical techniques in the environmental sciences, the new field of *environmetrics* has evolved into a promising new discipline. Moreover, many journals and books in which environmetrics research is published are mentioned in Section 1.6.3. In this section, various types of recently designed multivariate models, many of which are currently under development, are now discussed. Because of the controversy surrounding the disaggregation model, this family of models is entertained first.

20.5.2 Disaggregation Models

A special class of multivariate models is the *disaggregation* family of models. This class allows one to break down a series for which there are longer time units separating values into a sequence of values separated by shorter time units. For instance, an annual series can be disaggregated into a monthly series. The major reason why Valencia and Schaake (1973) proposed the disaggregation model was to insure that relevant statistics at both the annual and seasonal levels are consistent with one another. Annual flows, for example, could be generated by a short or long memory model and these annual flows could then be disaggregated to the seasonal level. As noted by Salas et al. (1985), disaggregation can be used for not only disaggregating variables in time, but for disaggregating in space as well. For example, precipitation over an area may be disaggregated into precipitation over sub-areas (Salas et al., 1980). Frevert and Lane (1986) presented a technique for accomplishing two level spatial disaggregation in a single run of their computer programs for disaggregation.

As discovered in a discussion with V. Klemes on May 30, 1985, in Tucson, Arizona, the basic idea of disaggregation is relatively old. In earlier research, the idea of disaggregation was utilized for addressing problems related to storage (Savarenskiy, 1940; Gould, 1961; Svanidze, 1962, 1980; Klemes, 1963, 1981). Woolhiser and Osborn (1986) devised a special kind of model for disaggregating storms into seasonal and regional components. Valencia and Schaake (1973) proposed a disaggregation model for obtaining seasonal flows from riverflows simulated at the annual level. As explained by Salas et al. (1985), since 1973 there have been numerous papers suggesting improvements to the original disaggregation model of Valencia and Schaake (1973) as well as related models developed thereafter. For example, Mejia and Rouselle (1976) put forward enhancements for the original disaggregation model of Valencia and Schaake (1973). Lee (1986) developed a multisite, multiseason synthetic flow generation model within a disaggregation framework. Other contributions to research in disaggregation are provided by authors including Tao and Delleur (1976), Stedinger and Vogel (1984), and Grygier and Stedinger (1988).

In a conversation held with V. Yevjevich on May 30, 1985, in Tucson, Arizona, he stated that two questions should be satisfactorily answered in order to adequately justify the use of disaggregation models in hydrology. The first question is whether there is information in annual measurements which is not contained in the seasonal observations. If there is not more information contained in the annual series, a better procedure may be to aggregate rather than disaggregate. When aggregating, a seasonal time series is modelled directly using a process such as a PARMA model and then it is aggregated to produce a compatible model for the aggregated series, usually at the annual level. Vecchia et al. (1983) presented a convincing argument which favours the concept of *aggregation* over disaggregation. They proved that if the original seasonal data follow a PARMA model in [14.2.15] or [14.2.16] with one moving average and one autoregressive parameter (i.e., PARMA(1,1)) then the aggregated annual data must be an ARMA model in [3.4.3] or [3.4.4] with one AR parameter (i.e., ARMA(1,0)) or else an ARMA model with one AR and one moving average parameter (i.e., ARMA(1,1)). Furthermore, there is significant gain in parameter estimation efficiency at the aggregated level when the seasonal data and their model are used rather than the aggregated (i.e., annual) data and their model. Rao et al. (1985) derived similar results for the situation where the seasonal data follow a PAR model in

[14.2.1] or [14.2.3]. In addition, they showed theoretically that the aggregated data can be more accurately predicted by using a valid model of the aggregated data. Moreover, in a related topic, aggregation of forecasts are discussed in Section 15.6 of this book. Further research regarding aggregation is presented by authors such as Kavvas et al. (1977), Obeysekera and Salas (1982, 1986) and Bartolini et al. (1988). Finally, Eagleson (1978) employs the principle of aggregation when he derives the distribution of annual precipitation from observed storm sequences.

A second issue raised by V. Yevjevich was whether the large number of parameters in a disaggregation model can be significantly reduced. In addition to other approaches, Stedinger et al. (1985) proposed a more parsimonious disaggregation model which is designed in a fashion which is analogous to the CARMA model described in detail in Chapter 21. Koutsoyiannis (1992) developed a multivariate dynamic disaggregation model, having a reduced parameter set, as a stepwise approach to disaggregation problems.

In certain situations, a practitioner or researcher may feel that it is appropriate to employ disaggregation models. A well-tested set of computer programs for implementing disaggregation models are available for use on both main frame and personal computers (Lane and Frevert, 1990).

20.5.3 Gaussian and NonGaussian Variables

As noted by Lewis (1985), simple linear models, such as the family of ARMA models, are not necessarily defined as having Gaussian variates but are simplest to use as such because linear operations on Gaussian variates preserves Gaussianity or normality. Furthermore, model construction procedures, based on the assumption of Gaussianity, are well developed. As a result, theoretical research regarding the development of stochastic models which can explicitly handle variables which are nonGaussian and therefore do not follow a normal distribution, has only been initiated recently. Among others, Tong et al. (1985) point out the fact that hydrological data are often not normally distributed and procedures are required to effectively handle this problem.

When the data are nonGaussian, one approach for obtaining data which are approximately normally distributed is to transform the original data using a transformation such as a Box-Cox transformation (Box and Cox, 1964) in [3.4.30]. This will produce a transformed series which is approximately Gaussian. A model based upon the Gaussian assumption can then be fitted to the transformed series. An alternative approach is not to assume Gaussianity in the first place but to select a distribution that the original data actually follow. Li and McLeod (1988) present results on estimation and diagnostic checking for ARMA models having nonGaussian innovations. Lewis (1985) describes a range of new models developed for use with continuous variate nonGaussian time series. The nonGaussian distributions he considers are the Exponential, Gamma, Weibull, Laplace, Beta and Mixed Exponential distributions. McKenzie (1985) presents a variety of models similar to Markov chains for describing discrete variate time series that follow various distributions The distributions which he entertains are the Poisson, Geometric, Negative Binomial and Binomial distributions. Finally, Brillinger (1985) develops procedures for fitting finite parameter models to nonGaussian series via bispectral fitting. The foregoing and other univariate developments in *nonGaussian modelling*, can be defined for modelling multiple time series. For example, Lewis (1985) mentions that his nonGaussian models can be extended to the multivariate case and that a periodic version of these models can be devised for modelling seasonal data.

20.5.4 Linear and Nonlinear Models

When a model is linear, it is a linear function of the variables in the model. In a *nonlinear model*, there is at least one term where the variables and/or innovations appear as products or are raised to powers. Lewis (1985) gives a brief discussion regarding alternative definitions for linearity and nonlinearity. Generally speaking, as the time interval between observations becomes smaller, the nonlinearities present in the data become more pronounced. For instance, average daily riverflow data may have to be modelled using a model containing nonlinear terms because of the nonlinear relationship between runoff and precipitation over a small time scale. On the other hand, a linear model may be sufficient to model mean annual riverflows for which the nonlinearities have been "averaged out".

A variety of interesting nonlinear stochastic models are now available (Tong, 1990). For example, Tong et al. (1985), Ozaki (1985), as well as Brillinger (1985) and Gallant (1987) describe threshold, discrete time storage, and nonlinear regression models, respectively, which are capable of modelling various kinds of nonlinearities which may be present in natural time series. Li (1992) derives the asymptotic standard errors of residual autocorrelations in nonlinear time series models for use in diagnostic checking while, in 1993, Li presents a statistical test for discriminating among different nonlinear time series models. As noted by Tong et al. (1985), the threshold model (Tong, 1983) can be easily defined for the multivariate case. Further, Brillinger (1985) defines a spatial-temporal process for multivariate modelling.

20.5.5 Multivariate Fractional Autoregressive-Moving Average (FARMA) Models

Short and long memory models are defined in Section 2.5.3 using [2.5.7]. A special class of long memory models is the FARMA family of models defined in [11.2.4]. Because the FARMA model is a generalized type of ARIMA (autoregressive integrated moving average) model given in [4.3.4] for which the differencing operator can assume real values, the FARMA model can be easily written for the multivariate case. However, a great deal of research is required to develop model construction techniques for use with multivariate FARMA models.

20.5.6 Time and Frequency Domains

Time series models such as the nonGaussian models of Lewis (1985) or the general multivariate ARMA model in [20.2.1] or [20.2.2] are defined in terms of discrete time variables. In order to fit a time series model to a data set, various techniques are available for use at the three stages of model construction. If a given method or statistic, such as the sample ACF which can be used for model identification, is expressed directly in terms of the time variable, it is said to be expressed in the *time domain*. Alternatively, one can work in the *frequency domain* by entertaining Fourier transforms. As explained in Sections 2.6 and 3.5, the Fourier transform of the autocovariance function produces the spectrum which expresses the distribution of the variance of the series with frequency (Jenkins and Watts, 1968). Although it is more common in water resources to execute univariate and multivariate time series modelling in the time domain rather than the frequency domain, sometimes it is advantageous to work in the latter domain. For instance, Brillinger (1985) presents interesting results regarding Fourier inference. Canfield and Bowles (1986) devise a method for conveniently generating multivariate series from the spectrum. Ghani and Metcalfe (1986) employ a spectral approach for predicting the probability of the peak flow exceeding a given level during a specified time period. A host of other applications of spectral methods to environmental problems can be found in journals and books referred to in 1.6.3, although only some of the published spectral research deals with multivariate problems.

20.5.7 Pattern Recognition

In order to model multivariate hydrological time series, MacInnes and Unny (1986) extend to the multivariate level the univariate approach of Panu and Unny (1980) and Unny et al. (1981) for modelling time series from a *pattern recognition* viewpoint. They apply their pattern recognition model to familiar multistation streamflow problems and discuss both the advantages and disadvantages of using pattern recognition-based models.

20.5.8 Nonparametric Tests

In order to lessen the number of underlying assumptions required for testing a hypothesis such as the presence of a specific kind of trend in a data set, researchers developed nonparametric procedures for use in hypothesis testing. Due to the great importance of nonparametric testing in environmental impact assessment, Chapter 23 of this book is entirely devoted to this type. Some of the nonparametric tests, such as the partial rank correlation tests of Section 23.3.6, can either be used or else extended for use with multi-site and/or multiple variable data sets. As a result, nonparametric tests are very useful in multivariate analysis, especially when the data are very *messy*. Part X of this book explains how intervention analysis, nonparametric tests and regression analysis can be used for modelling *messy environmental data*.

20.6 CONCLUSIONS

The general multivariate ARMA model is defined in [20.2.1] and [20.2.2] while model construction procedures are presented in Section 20.3.2 and Appendix A20.1. Within this general family of time series models, the CARMA and TFN models are of particular importance in the field of water resources and environmental engineering. As explained in this chapter and Section 21.1 and also by Salas et al. (1985) and Camacho et al. (1985a, 1986), the physical constraints dictated by a given hydrological system negate the need for using the general form of the multivariate ARMA model and, therefore, usually some type of CARMA or TFN model is all that is required in a practical application. Additionally, as described in depth in Chapters 17 and 21, model construction techniques are now fully developed for employment with TFN and CARMA models, respectively. The CARMA model constitutes a parsimonious version of the general multivariate ARMA model for describing multiple time series that are contemporaneously correlated with one another. Besides being able to model the impacts of interventions upon the mean level of a series and estimate missing observations (see Chapter 19), the TFN model can describe the mathematical relationships between a single response variable and any number of covariate series (see Chapter 17). The most general form of the TFN or intervention model is defined in [19.5.8].

In Section 20.4 the historical development of multivariate ARMA modelling in hydrology is outlined while other kinds of families of multivariate models are referred to in Section 20.5. These additional types of multivariate models include various classes of disaggregation, nonGaussian, nonlinear and long memory models. For some of these models, such as the nonGaussian models of Lewis (1985) and the nonlinear threshold models of Tong et al. (1985), further research is required for developing model construction techniques for use in modelling multiple time series.

For classifying the capabilities of a family of time series models, Hipel (1985b) suggests a list of twenty-five criteria that reflect the main statistical characteristics that could be modelled. This list includes linear, nonlinear Gaussian, nonGaussian, long memory and short memory criteria. Using these criteria, a given multivariate model, such as the CARMA model, can be categorized as being linear, Gaussian and short memory. By understanding the modelling capabilities of each family of models, as well as the main physical and statistical characteristics of the time series being studied, a practitioner can decide upon which classes of models are most appropriate to consider for modelling the key statistical properties of his or her data set. Subsequent to exploratory data analysis referred to in Section 1.2.4 and described in detail in Section 22.3, the practitioner can select the best specific time series model at the confirmatory data analysis stage by following the three stages of model construction.

An obvious extension to the work completed thus far in time series analysis is to develop more comprehensive families of models that can simultaneously handle a wider variety of the criteria. For example, it may be possible to design a multivariate model that can take care of both nonlinear and nonGaussian characteristics of the data. However, any new class of models should be designed to be as simple as possible and thereby not have too many parameters, as well as provide a good statistical fit to the data. Whenever possible, researchers are encouraged to incorporate both the physical and statistical aspects of the problem into the basic model design. Subsequent to the design, appropriate algorithms are required for use at the three stages of model construction. A continuous dialogue among the statisticians, water resources engineers and other scientists should increase the probability of designing new models that will be welcomed by the practitioners for solving pressing water resources problems.

Besides designing new models, existing models should be rigorously compared from both theoretical and empirical viewpoints in order to ascertain which families of models are most appropriate to use in practice. For instance, thorough scientific comparisons of the disaggregation and aggregation approaches to time series modelling are long overdue. "To disaggregate or not to disaggregate", that is the nagging question haunting both practitioners and theoreticians alike in hydrology.

Due to the continued and growing abuse of the natural environment by man-induced activities such as industrialization and agricultural development, there will continue to be a great demand for having flexible multivariate models for use in environmental impact assessment. Future research in the time series aspects of environmental impact assessment will probably entail developing more nonparametric tests for handling a wider variety of situations in trend detection and evaluation, rigorously comparing the capabilities of both parametric and nonparametric approaches, and providing guidelines for optimally designing sampling schemes for water quality variables.

As is also emphasized in Section 3.6 and elsewhere in the book, Yevjevich and Harmancioglu (1985) stress the importance of linking stochastic models with physical consistent properties of any particular water resources time series. In some circumstances, it may be possible to employ purely stochastic models to accomplish this goal. Alternatively, it may be necessary to employ a combination of deterministic and stochastic models for realistically modelling certain kinds of water resources systems.

The authors maintain that the development and use of multivariate models, in general, and multivariate time series models in particular, will significantly increase in the future within the realm of water resources and environmental engineering. Besides hydrological time series, multivariate models will be used more and more for modelling other kinds of water related time series such as water quality (physical, chemical and biological), water demand, water pricing and meteorological time series.

APPENDIX A20.1

IDENTIFICATION METHODS FOR GENERAL MULTIVARIATE ARMA MODELS

As advocated by Tiao and Box (1981), the sample CCF and PACF matrices are especially useful for identifying pure multivariate MA and AR models, respectively. When the set of time series follow a multivariate ARMA(p,q) process, the extended sample cross correlation function (ESCCF) matrix of Tsay and Tiao (1984) and Tiao and Tsay (1983a,b) can also be used. Section 20.3 describes how model construction is carried out for general multivariate ARMA models.

A20.1.1 Sample CCF Matrix

Suppose k time series of length n are represented at time t by the vector

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$$

For lag l, where $l = 1,2,\ldots,$ the *theoretical CCF matrix* of order $k \times k$ is written as $\rho(l)$. A typical entry, $\rho_{ij}(l)$, in the matrix is theoretically defined as

$$\rho_{ij}(l) = Cov(Z_{ti}, Z_{t+l,j})/[Var(Z_{ti}) \cdot Var(Z_{tj})]^{1/2} \qquad [A20.1.1]$$

The *sample CCF matrix* at lag l is denoted by $\mathbf{R}(l)$. Each element $r_{ij}(l)$ in the matrix is calculated using

$$r_{ij}(l) = \sum_{t=1}^{n-l}(Z_{ti} - \bar{Z}_i)(Z_{t+l,j} - \bar{Z}_j)/[\sum_{t=1}^{n}(Z_{ti} - \bar{Z}_i)^2 \sum_{t=1}^{n}(Z_{tj} - \bar{Z}_j)^2]^{1/2} \qquad [A20.1.2]$$

where $\bar{Z}_i$ and $\bar{Z}_j$ are the sample means of the ith and jth series, respectively.

For a pure multivariate MA process of order q, the theoretical CCF matrix vanishes after lag q (Tiao and Box, 1981; Jenkins and Alavi, 1981). Hence, if $\mathbf{Z}_t$ follows a MA(q) process, the entries in the sample CCF matrix are not significantly different from zero for $l > q$. Because the asymptotic distribution of $r_{ij}l$ is N(0,1/n), the 95% confidence limits given approximately by $\pm\frac{2}{n^{1/2}}$ can be used to decide whether or not the estimated value is significantly different from zero. If each entry in $\mathbf{R}(l)$ falls within these limits, it can be assumed that the sample CCF matrix has cutoff at lag l and, consequently, l can be considered as the order of the multivariate MA model. If the series $\mathbf{Z}_t$ follows a multivariate AR model of order p or a general multivariate

ARMA(p,q) model, no cut-off will appear in $\mathbf{R}(l)$. Tiao and Box (1981) suggest summarizing the numerical values of each $r_{ij}(l)$ with "+" to indicate a value greater than $\frac{2}{n^{1/2}}$, with "-" to indicate a value less than $\frac{-2}{n^{1/2}}$ and with "." to indicate a value between $\frac{-2}{n^{1/2}}$ and $\frac{2}{n^{1/2}}$. This is a very convenient way to interpret the entries in $\mathbf{R}(l)$. Similar to the situation for the identification of univariate ARMA models in Chapters 3 and 5, one must define another statistic for detecting cutoff in pure multivariate AR(p) models.

A20.1.2 Sample PACF Matrix

For a pure multivariate AR process of order p, denoted by AR(p), the theoretical PACF matrix, $\mathbf{P}(l)$, for lag l, where $l = 1,2,\ldots,$ can be defined. Let Φ_{lj} be the jth AR matrix in a multivariate AR process of order l so that Φ_{ll} is the last matrix. The *theoretical PACF matrix*, $\mathbf{P}(l)$, is defined as $\Phi_{l,l}$ where $\Phi_{l1},\Phi_{l2},\ldots,\Phi_{ll}$, are the solutions of the system of equations

$$\sum_{i=1}^{l}\Phi_{li}\cdot\Gamma(j-i)=\Gamma(-j),\ j=1,2,\ldots,l \qquad [A20.1.3]$$

where $\Gamma(j)=Cov[\mathbf{Z}_t,\mathbf{Z}_{t+j}]$. This equation constitutes a multivariate generalization of the Yule-Walker equations in [3.2.17] for AR models in univariate time series analysis. If the process is multivariate AR(p) then by definition $\Phi_{pp}=\Phi_p$ in [20.2.1] and $\Phi_{p+j,p+j}=\mathbf{0}$ for $j>0$. Consequently, for a pure multivariate AR(p) process, $\mathbf{P}(l)=\mathbf{0}$ for $l>p$. If the process is MA(q) or ARMA(p,q), Φ_{ll} will not cutoff but rather decay to zero.

The *sample PACF matrix* at lag l where $l = 1,2,\ldots,$ is defined as the $\hat{\mathbf{P}}(l)=\hat{\Phi}_{ll}$ matrix in the solution of the system of equations

$$\sum_{i=1}^{l}\hat{\Phi}_{li}\mathbf{R}(j-i)=\mathbf{R}(-j) \qquad [A20.1.4]$$

The sample PACF matrix, $\hat{\mathbf{P}}(l)$ at lag l, is calculated by fitting a multivariate AR(l) model using

$$\mathbf{Z}_t=\mathbf{c}+\Phi_{l1}\mathbf{Z}_{t-1}+\Phi_{l2}Z_{t-2}+\cdots+\Phi_{l,l}\mathbf{Z}_{t-l}+\mathbf{a}_t^{(l)} \qquad [A20.1.5]$$

and setting $\hat{\mathbf{P}}(l)=\Phi_{ll}$, where $\mathbf{c}$ is a vector of k constants that are recursively estimated along with the other model parameters using standard multivariate least squares (Tiao and Box, 1981). Asymptotically, the distribution for each entry in $\hat{\mathbf{P}}(l)$ is $N(0,\frac{1}{n})$. To denote whether an entry in $\hat{\mathbf{P}}(l)$ is greater than, less than, or falls within the approximate 95% confidence limits given by $\pm\frac{2}{n^{1/2}}$, one can employ "+", "-" or ".", respectively. Using the foregoing symbols rather than numerical values makes it easier to detect the important identification information contained in $\hat{\mathbf{P}}(l)$.

Basically, the sample PACF matrices are determined by fitting AR models of order $l = 1,2,\ldots,$ in [A20.1.5]. By analyzing the variance-covariance matrices corresponding to successive AR fittings, one can ascertain how much the statistical fit improves as the order l is increased.

For any given l, $l = 1,2,\ldots,$ a formal test of hypothesis for testing the null hypothesis: **P**(l)=**0** against the alternative $\mathbf{P}(l) \neq \mathbf{0}$ can be performed. The likelihood ratio statistic is given by

$$U = |Sl|/|S(l-1)| \qquad \text{[A20.1.6]}$$

where

$$S(l) = \sum_{t=1}^{n} \hat{\mathbf{a}}_t^{(l)} \cdot \hat{\mathbf{a}}_t^{(l)T}$$

for which $\hat{\mathbf{a}}_t^{(l)}$ are the residuals of the fitted model in [A20.1.5].

Using Barlett's (1938) approximation, the statistic

$$\chi(l) = -(N - 1/2 - lk) \cdot log[|S(l)|/|S(l-1)|] \qquad \text{[A20.1.7]}$$

will have a chi-square distribution with k^2 degrees of freedom when $l > p$. Now, for MA(q) or general ARMA(p,q) models, the sample PACF matrices do not have a cutoff and, therefore, they are expected to obtain significant values of $\chi(l)$ even for large lags. Examples of the use of the sample ACF and PACF matrices in the identification of hydrologic time series are given by Camacho et al. (1987c).

The sample ACF and PACF matrices are very useful for identifying pure MA and pure AR models, respectively. In practice, the difficulty arises in the identification of mixed ARMA(p,q) models when both p and q are larger than zero. In these cases, the ESCCF can be employed to help in the identification.

A20.1.3 ESCCF Matrix

The *ESCC (extended sample cross correlation) matrix* was proposed by Tsay and Tiao (1984) and Tiao and Tsay (1983a,b) to help in the identification of the order of mixed multivariate ARMA models. The main idea of the technique is to calculate consistent estimates of the AR matrices $\Phi_1, \ldots, \Phi_p$, say, and then use the properties of the transformed series

$$\mathbf{W}_t = \mathbf{Z}_t - \sum_{l=1}^{p} \hat{\Phi}_l \mathbf{Z}_{t-l} \qquad \text{[A20.1.8]}$$

to identify the order of the process. If $\mathbf{Z}_t$ follows a multivariate ARMA(p,q) process and the estimated parameters are consistent, then $\mathbf{W}_t$ follow a multivariate MA(q) process, so that the sample CCF matrices of $\mathbf{W}_t$ should be able to identify the proper order of the model. Tiao and Tsay (1983a) have shown that when the order of the model is known, it is possible to obtain consistent estimates for the Φ's using a process of iterated regressions. However, in practice the order of the model is not known in advance, and therefore, it is necessary to study the properties of the iterated regression estimates and their associated transformed series $\mathbf{W}_t$.

The following algorithm can be used to obtain the iterated regression estimates where further details are given by Tiao and Tsay, (1983a):

Step 1: For $m = 1,2,\ldots,M{+}J{+}1$ fit an AR(m) regression

$$\mathbf{Z}_t = \sum_{l=1}^{m} \Phi_{l(m)}^{(o)} \mathbf{Z}_{t-l} + emt^{(o)}, \quad t = m+1, \ldots, n \qquad [A20.1.9]$$

using ordinary least squares. Denote the estimated parameters as $\hat{\Phi}_{l(m)}^{(o)}$. The superscript (o) indicates the ordinary least square estimates and the subscript (m) indicates the order of the AR fit.

Step 2: For $j = 1,2, \ldots, J$, and $m = 1, \ldots, M$, recursively compute the AR(m) jth iterated estimates $\hat{\Phi}_{l(m)}^{(j)}$, $l = 1, \ldots, m$ as:

$$\hat{\Phi}_{l(m)}^{(j)} = \hat{\Phi}_{l(m+1)}^{(j-1)} - \hat{\Phi}_{m+1(m+1)}^{(j-1)} [\hat{\Phi}_{m(m)}^{(j-1)}]^{-1} \hat{\Phi}_{l-1(m)}^{(j-1)} \qquad [A20.1.10]$$

where $\hat{\Phi}_{o(m)}^{(j-1)} = -\mathbf{I}$.

Now if $\mathbf{Z}_t$ follows a multivariate ARMA(p,q) and no linear combination of $\mathbf{Y}_{pt} = (\mathbf{Z}_t^T, \ldots, \mathbf{Z}_{t-p+1})^T$ follows a MA model with order less than q, then the matrices $\hat{\Phi}_{l(m)}^{(j)}$ are consistent estimators for Φ_l when (i) $m \geq p$ and $j = q$, or, (ii) $m = p$ and $j > q$ (Tiao and Tsay, 1983a). Also, for $j > q$, the transformed series

$$\mathbf{W}_{p,t}^{(j)} = \mathbf{Z}_t - \sum_{l=1}^{m} \hat{\Phi}_{l(p)}^{(j)} \mathbf{Z}_{t-l} \qquad [A20.1.11]$$

will approximately follow a multivariate MA(q) model. Therefore, the sample CCF matrix of $W_{p,t}^{(j)}$ will have a cutoff after lag q. In particular, the lag j CCF matrix of $\mathbf{W}_{p,t}^{(j)}$, $\hat{\mathbf{\rho}}_{(p)}(j)$ will be such that

$$\hat{\rho}_{(p)}(j) \xrightarrow{p} \begin{cases} \mathrm{C}, & j \leq q \\ \mathbf{0}, & j > q \end{cases} \qquad [A20.1.12]$$

where C is a generic symbol for a matrix whose elements are not necessarily all zero.

The mth ESCCF matrix is now defined as $\hat{\rho}_{(m)}(j)$ is the lag j sample CCF matrix of $\mathbf{W}_{m,t}^{(j)}$. For a general multivariate ARMA(p,q) process, $\hat{\rho}_m(j)$ has the following asymptotic property:

$$\hat{\rho}_{(m)}(j) \xrightarrow{p} \begin{cases} \mathbf{0}, & 0 \leq m-p < j-q \\ \mathrm{C}, & \text{otherwise} \end{cases} \qquad [A20.1.13]$$

This property of the ESCCF can now be exploited in the following way to help in the identification of the order (p,q) of an ARMA model:

Stage 1: Arrange the ESCCF matrices $\hat{\rho}_{(m)}(j)$ in a block matrix as is shown in Table A20.1.1. The rows numbered 0,1,2, . . . , signify the AR order and, similarly, the columns signify the MA order. To investigate how to use this table, suppose that the true model for $\mathbf{Z}_t$ is a multivariate ARMA(2,1) model. The pattern of the asymptotic behaviour of the ESCCF is shown in Table A20.1.2.

It can be observed from Table A20.1.2 that there is a triangle of asymptotic zero matrices and that the vertex of the triangle is located at the entry (2,1). For a general multivariate ARMA(p,q) model the same pattern is expected and the vertex of such a triangle will always be located at entry (p,q). This observation suggests the second idea in the identification of the process.

Table A20.1.1. The ESCCF table.

AR	MA					
	0	1	2	3	·	·
0	$\hat{\rho}_{1(0)}$	$\hat{\rho}_{2(0)}$	$\hat{\rho}_{3(0)}$	$\hat{\rho}_{4(0)}$	·	·
1	$\hat{\rho}_{1(1)}$	$\hat{\rho}_{2(1)}$	$\hat{\rho}_{3(1)}$	$\hat{\rho}_{4(1)}$	·	·
2	$\hat{\rho}_{1(2)}$	$\hat{\rho}_{2(2)}$	$\hat{\rho}_{3(2)}$	$\hat{\rho}_{4(2)}$	·	·
3	$\hat{\rho}_{1(3)}$	$\hat{\rho}_{2(3)}$	$\hat{\rho}_{3(3)}$	$\hat{\rho}_{4(3)}$	·	·
·	·	·	·	·	·	·
·	·	·	·	·	·	·

Table 20.1.2 The asymptotic ESCCF matrix table of an ARMA(2,1) model where C and 0 denote, respectively, a nonzero and a zero matrix.

AR	MA					
	0	1	2	3	4	5
0	C	C	C	C	C	C
1	C	C	C	C	C	C
2	C	0	0	0	0	0
3	C	C	0	0	0	0
4	C	C	C	0	0	0

Stage 2: Look for the vertex with entry (d_1,d_2) of a triangle of asymptotic zero matrices with boundary lines $m = d_1$ and $j-m = d_2 \geq 0$ and tentatively specify the order of the model as $p = d_1$ and $q = d_2$. As a crude approximation, the value $(n-m-j)^{-1}$ can be used for the variance of the elements of $\hat{\rho}_{(m)}(j)$ under the hypothesis that the transformed series $\mathbf{W}_{m,j}^{(j)}$ is a white noise process. As an informative summary, (+, ·, -) signs can be used to replace numerical values of the elements of $\hat{\rho}_{m(j)}$, where a plus sign (+) is used to indicate a value greater than two times the standard deviation, a minus sign (-) to indicate a value less than minus two times the standard deviation, and a period (·) for in-between values.

Simulated and hydrological applications of using the ESCCF for multivariarte ARMA(p,q) model identification are given by Camacho et al. (1986). Consider, in particular, the simulated example. On hundred realizations of an ARMA(1,1) model with parameter matrices

$$\Phi = \begin{bmatrix} .8 & .0 \\ .5 & .7 \end{bmatrix} \quad \Theta = \begin{bmatrix} .5 & .0 \\ .0 & .5 \end{bmatrix} \quad \Delta = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix} \qquad \text{[A20.1.14]}$$

were generated using the simulation algorithm of Camacho (1984) which is similar to the one described in Section 21.4.2 for CARMA models. The plot of the data is given in Figure A20.1.1 and the pattern of the ESCCF matrices is shown in Table A20.1.3. As can be observed, the vertex of a triangle of zero matrices is located at entry (1,1), indicating that an ARMA(1,1) model would be adequate to fit the data.

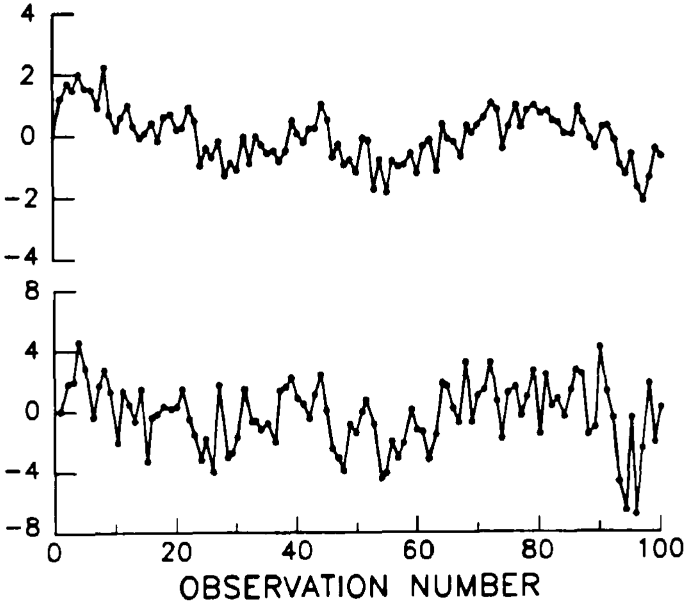


Figure A20.1.1. 100 simulated observations of an ARMA(1,1) model with parameter matrices given in [A20.1.14].

Table A20.1.3. ESCCF matrix for the simulated ARMA(1,1) data.

	MA					
AR	0	1	2	3	4	5
0	+ + + +	+ + + +	• + + +	• + • +	• + • +	• • • +
1	+ + • •	• • • •	• • • •	• • • •	• • • •	• • • •
2	+ − + −	• + • •	• • • •	• • • •	• • • •	• • • •
3	• − • •	− − • •	• − • −	• • • •	• • • •	• • • •
4	• + • +	• • • •	+ • + •	• • − •	• • • •	• • • •

PROBLEMS

20.1 The general multivariate ARMA model is defined in [20.2.1] and [20.2.2]. Write down the following vector ARMA(p,q) models using both the matrix notation and the full length description when matrices and vectors are not used.

(a) ARMA(4,0)

(b) ARMA(0,3)

(c) ARMA(2,2)

(d) ARMA(3,1)

20.2 Demonstrate that the two equations for writing a TFN model given in [20.2.3] and [17.5.3] are equivalent.

20.3 Explain why the PARMA model given in [14.2.15] can be considered as a special case of the multivariate ARMA model in [20.2.1].

20.4 By referring to [20.2.1] give the matrix equations as well as the equations where matrices and vectors are not employed for the following CARMA(p,q) models:

(a) CARMA(4,0)

(b) CARMA(0,3)

(c) CARMA(2,2)

(d) CARMA(3,1)

20.5 The general multivariate ARMA model for fitting to a set of seasonal time series is defined in [20.2.1] and [20.2.2]. Assuming that there are 5 seasons per year, define the general multivariate deseasonalized ARMA model using both the matrix notation and the full length notation when matrices and vectors are not used. Explain how you would fit this model to a set of seasonal time series by following the three stages of model construction. In your explanation be sure to mention specific graphical methods and algorithms that you would employ.

20.6 Repeat the instructions of problem 20.5 for the case of a general multivariate PAR model.

20.7 Follow the instructions of problem 20.5 for the case of a general multivariate PARMA model.

20.8 In appendix A20.1, three procedures are presented for identifying general multivariate ARMA models. By referring to this appendix and also the original references, compare the advantages and drawbacks of these approaches. Explain how they could be expanded for use with seasonal data.

20.9 Using equations when necessary, outline the approach of Hillmer and Tiao (1979) for estimating the parameters of a general multivariate ARMA(p,q) model.

20.10 Employing equations when needed, summarize the procedure of Ansley and Kohn (1983) for calibrating a general multivariate ARMA model.

20.11 Explain how the sample CCF and sample PACF described in Appendix A20.1 can be employed for checking the whiteness assumption about the residuals of a calibrated general multivariate ARMA model.

20.12 Define the modified Portmanteau test of Li and McLeod (1981) and explain how it can be utilized for testing the whiteness assumption of the residuals of a fitted general multivariate ARMA model.

20.13 Using equations when necessary, explain how multivariate normality tests proposed by Royston (1983) can be employed for testing the normality assumption for the residuals of a calibrated general multivariate ARMA model.

20.14 Select two nonseasonal time series which you suspect should be modelled using some type of multivariate ARMA model. Using the residual CCF approach presented in detail in Section 16.3.2 and outlined in Section 20.3.2, determine what kind of multivariate ARMA model could be fitted to this data.

20.15 For the two nonseasonal time series examined in problem 20.14, calibrate the most appropriate multivariate ARMA(p,q) model.

20.16 Execute the instructions of problem 20.14 for the case of two seasonal time series.

20.17 Follow the instructions of problem 20.15 for the seasonal time series examined in problem 20.16.

20.18 Mathematically define the input-output class of models put forward by Cooper and Wood (1982a,b). Explain the assets and drawbacks of this group of models, especially with respect to the general multivariate ARMA family of models. How are the input-output models mathematically related to the multivariate ARMA models?

20.19 The STARMA (space-time ARMA) class of models is discussed in Section 20.4. By referring to appropriate references, mathematically define this group of models, summarize its advantages and drawbacks, and compare it to the multivariate ARMA family of models in [20.2.1] and [20.2.2].

20.20 Define mathematically the multivariate FGN model of Matalas and Wallis (1971) referred to in Section 20.4. Is this a realistic model to employ in practice? Justify your response. You may wish to refer to the discussions on FGN given in Section 10.5 as well as appropriate references listed at the end of Chapter 10.

20.21 Within the hydrological literature there has been an ongoing and heated debate about whether one should employ disaggregation or aggregation models in hydrology. By referring to appropriate research work that is referenced in Section 20.5.2, mathematically define a disaggregation model and also an aggregation model based upon ARMA processes. Compare these two categories of models according to their relative advantages and disadvantages. Which class of models would you employ in hydrological applications?

20.22 Using mathematical equations, extend the nonGaussian model of Lewis (1985) mentioned in Section 20.5.3 to the multivariate case. Discuss any implementation problems that you may encounter when applying this mathematical model.

20.23 Define mathematically the multivariate version of the threshold model of Tong et al. (1985) referred to in Section 20.5.4. Describe the types of data to which you think this model could be applied and discuss potential model construction techniques.

20.24 Mathematically define the multivariate FARMA model of Section 20.5.5. Describe the types of model construction tools that would have to be developed to apply this model in practice.

REFERENCES

CARMA MODELS

Camacho, F. (1984). Contemporaneous CARMA Modelling with Applications. Ph.D. thesis, Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1985a). Contemporaneous autoregressive-moving average (CARMA) modelling in hydrology. *Water Resources Bulletin*, 21(4):709-720.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1986). (July 15-17, 1985). Developments in multivariate ARMA modelling in hydrology. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, Fort Collins, Colorado. Engineering Research Center, Colorado State University, pages 178-197.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987a). Contemporaneous bivariate time series. *Biometrika*, 74(1):103-113.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987b). Multivariate contemporaneous ARMA models with hydrological applications. *Stochastic Hydrology and Hydraulics*, 1:141-154.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987c). The use and abuse of multivariate time series models in hydrology. In MacNeill, I. B. and Umphrey, G. J., editors, *Advances in the Statistical Sciences*, Festschrift in Honor of Prof. V. M. Joshi's 70th Birthday, Volume IV, Stochastic Hydrology, pages 27-44. D. Reidel, Dordrecht, The Netherlands.

Chan, W. T. and Wallis, K. F. (1978). Multiple time series modelling: Another look at the mink-muskrat interaction. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 27(2):168-175.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press, New York.

Hannan, E. J. (1970). *Multiple Time Series*. John Wiley, New York.

Risager, F. (1980). Simple correlated autoregressive process. *Scandinavian Journal of Statistics*, 7:49-60.

Risager, F. (1981). Model checking of simple correlated autoregressive processes. *Scandinavian Journal of Statistics*, 8:137-153.

Wallis, K. F. (1977). Multiple time series analysis and the final form of econometric models. *Econometrica*, 45(6):1481-1497.

CONCEPTUAL MODELS

Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985). Forecasting quarter-monthly riverflow. *Water Resources Bulletin*, 21(5):731-741.

DISAGGREGATION AND AGGREGATION

Eagleson, P. S. (1978). Climate, soil, and vegetation, 2. The distribution of annual precipitation derived from observed storm sequences. *Water Resources Research*, 14(5):713-721.

Frevert, D. K. and Lane, W. L. (1986). Development and application of a two level spatial disaggregation procedure for the LAST statistical hydrology package. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, July 15-17, 1985, Fort Collins, Colorado. Engineering Research Center, Colorado State University, pages 198-216.

Gould, B. W. (1961). Statistical methods for estimating the design capacity of dams. *Journal of the Institute of Engineers in Australia*, 33(12):405-416.

Grygier, J. C. and Stedinger, J. R. (1988). Condensed disaggregation procedures and conservation corrections for stochastic hydrology. Water Resources Research, 24(10):1574-1584.

Kavvas, M. L., Cote, L. J., and Delleur, J. W. (1977). Time resolution of the hydrologic timeseries models. *Journal of Hydrology*, 32:347-361.

Klemes, V. (1963). Seasonal component of storage in statistical methods of streamflow control (in Czechoslovakian). *Vodohospid. Cas.*, 11(3):232-247, 11(4):341-360.

Klemes, V. (1981). Applied stochastic theory of storage in evolution. *Advances in Hydrosciences*, Academic Press, New York, 12:79-141.

Koutsoyiannis, D. (1992). A nonlinear disaggregation method with a reduced parameter set for simulation of hydrologic series. *Water Resources Research*, 28(12):3175-3191.

Lane, W. L. and Frevert, D. K. (1990). *Applied Stochastic Techniques, User Manual, LAST Personal Computer Package Version 5.2*. United States Department of the Interior, Bureau of Reclamation, Denver Office, P. O. Box 25007, Denver Federal Center, Denver, Colorado 80225-0007, Tel (303) 236-3809.

Lee, H.-L. (1986). Preserving cross-year correlations among seasonal flow series using disaggregation procedures. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, July 15-17, 1985, Colorado State University, Fort Collins, Colorado, pages 595-604.

Mejia, J. M. and Rousselle, J. (1976). Disaggregation models in hydrology revisited. Water Resources Research, 12(2):185-186.

Obeysekera, J. T. B. and Salas, J. D. (1982). On the aggregation and disaggregation of streamflow time series. *American Geophysical Union EOS*, 63(18):321.

Obeysekera, J. T. B. and Salas, J. D. (1986). Modeling of aggregated hydrologic time series. *Journal of Hydrology*, 86:197-219.

Rao, A. R., Rao, S. G., and Kashyap, R. L. (1985). Stochastic analysis of time-aggregated hydrologic data. *Water Resources Bulletin*, 21(5):757-771.

Savarenskiy, A. D. (1940). A method for streamflow control computation. *Gidrotekhn. Stroit.*, 2:24-28. first released in 1938 as a report of Kuybyshevskiy Stroitelnyy Instit., Kuybyshev, (in Russian).

Stedinger, J. R., Lettenmaier, D. P., and Vogel, R. M. (1985). Multisite ARMA(1,1) and disaggregation models for annual streamflow generation. *Water Resources Research*, 21(4):497-507.

Stedinger, J. R. and Vogel, R. M. (1984). Disaggregation procedures for generating serially correlated flow vectors. *Water Resources Research*, 20(1):47-56.

Svanidze, G. G. (1962). Simulation of hydrologic series with regard for intra-annual distribution of runoff (fragments method). *Dokl. Gidrol. Konf. CHSSR* (Reports of the Hydrological Conference of the Czechoslovak Socialist Republic), pages 411-418. (See also Tr. In-ta Energetiki AN GSSR, Vol. 17, 1963, pages 273-282, and VODOHOSPODARSKY CHASOPIS, No. 6, 1963, pages 138-149.).

Svanidze, G. G. (1980). *Mathematical Modeling of Hydrologic Series for Hydroelectric and Water Resources Computations*. Water Resources Publications, Littleton, Colorado. (Translated from the Russian edition, Gidrometeoizaat, Leningrad, USSR, 1977.).

Tao, P. C. and Delleur, J. W. (1976). Multistation, multiyear synthesis of hydrologic time series by disaggregation. *Water Resources Research*, 12(6):1303-1312.

Valencia, D. R. and Schaake, J. C. (1973). Disaggregation processes in stochastic hydrology. *Water Resources Research*, 9(3):580-585.

Woolhiser, D. A. and Osborn, H. B. (1986). Point storm disaggregation - seasonal and regional effects. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes*, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes, July 15-17, 1985, Fort Collins, Colorado. Colorado State University, pages 105-120.

LONG MEMORY MODELS

Franchini, M., Todini, E., Giuliano, G., and O'Connell, P. E. (1986). A multivariate multiseasonal model for long term persistent hydrological series. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes*, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes, July 15-17, 1985, Fort Collins, Colorado. Colorado State University, pages 243-262.

Matalas, N. C., and Wallis, J. R. (1971). Statistical properties of multivariate fractional noise processes. *Water Resources Research*, 7(6):1460-1468.

O'Connell, P. E. (1974). *Stochastic Modelling of Long-Term Persistence in Stream Flow Sequences.* Ph.D. Thesis, Civil Engineering Dept., Imperial College, London.

MULTIVARIATE ARMA MODELS

Abraham, B. (1980). Intervention analysis and multiple time series. *Biometrika,* 67(1):73-78.

Ansley, C. F. and Kohn, R. (1983). Autoregressive-moving average process with missing or aggregated data. *Biometrika,* 70:275-278.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society,* 34:33-40.

Bernier, J. (1971). Modeles probabilistes a variables hydrologiques multiples et hydrologie synthetique. *Mathematical Models in Hydrology,* IAHS-AISH Publication No. 100.

Canfield, R. V. and Tseng, L. (September 1979). Homogeneous ARMA processes. Paper presented at the American Water Resources Association Conference, Las Vegas, Nevada.

Cooper, D. M. and Wood, E. F. (1982a). Identification of multivariate time series and multivariate input-output models. *Water Resources Research,* 18(4):937-946.

Cooper, D. M. and Wood, E. F. (1982b). Parameter estimation of multiple input-output time series models: Application to rainfall-runoff processes. *Water Resources Research,* 18(5):13 52-1364.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B, 39:1-38.

Fiering, M. B. (1964). Multivariate techniques for synthetic hydrology. *Journal of the Hydraulics Division,* ASCE, 90(HY5):43-60.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica,* 37(3):424-438.

Hall, A. D. and Nicholls, D. F. (1980). The evaluation of exact maximum likelihood estimates for VARMA models. *Journal of Statistical Computation and Simulation,* 10:251-262.

Hannan, E. J. and Kavalieris, L. (1984). Multivariate linear time series models. *Advances in Applied Probability,* 16:492-561.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society,* Series B, 41(2):190-195.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika,* 69:81-94.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association,* 74(367):652-660.

Hipel, K. W. (1985a). Time series analysis in perspective. *Water Resources Bulletin,* 21(4)609-624.

Hipel, K. W., editor (1985b). *Time Series Analysis in Water Resources.* The American Water Resources Association, Bethesda, Maryland.

Hipel, K. W. (1986). Stochastic research in multivariate analysis. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, Keynote Address, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes,* July 15-17, 1985, Fort Collins, Colorado. Engineering Research Center, Colorado State University, pages 2-50.

Jenkins, G. M. and Alavi, A. S. (1981). Some aspects of modeling and forecasting multivariate time series. *Journal of Time Series Analysis,* 2(1):1-47.

Kahan, J. P. (1974). A method for maintaining cross and serial correlations and the coefficient of skewness under generation in a linear bivariate regression model. *Water Resources Research,* 10:1245-1248.

Kuczera, G. (1987). On maximum likelihood estimators for the multisite lag-one streamflow model: Complete and incomplete data cases. *Water Resources Research,* 23(4):641-645.

Lawrance, A. J. (1976). A reconsideration of the fiering two-station model. *Journal of Hydrology,* 29:77-85.

Ledolter, J. (1978). The analysis of multivariate time series applied to problems in hydrology. *Journal of Hydrology,* 36:327-352.

Lettenmaier, D. P. (1980). Parameter estimation of multivariate streamflow synthesis. In *Proceedings of the Joint Automatic Control Conference,* August 13-15, 1980, paper number FA6-D, San Francisco.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society,* Series B, 43(2):231-239.

Maas, A., Hufschmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A., and Fair, M. G. (1962). *Design of Water-Resource Systems.* Harvard University Press, Cambridge, Massachusetts.

Matalas, N. C. (1967). Mathematical assessment of synthetic hydrology. *Water Resources Research,* 3(4):937-945.

Newbold, P. and Hotopp, S. M. (1984). Multiple time series model building and testing for causality. Working paper, Department of Economics, University of Illinois, Urbana-Champaign.

Nicholls, D. F. and Hall, A. D. (1979). The exact likelihood of multivariate autoregressive-moving average models. *Biometrika,* 66:259-264.

Pierce, D. A. and Haugh, L. D. (1977). Causality in temporal systems. *Journal of Econometrics,* 5:265-293.

Quinn, B. G. (1980). Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society,* Series B, 42:182-185.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 32(2):121-133.

Salas, J. D., Delleur, J. W., Yevjevich, V., and Lane, W. L. (1980). *Applied Modelling of Hydrologic Time Series.* Water Resources Publications, Littleton, Colorado.

Salas, J. D. and Pegram, G. G. S. (1978). A seasonal multivariate multilag autoregressive model in hydrology. In Morel-Seytoux, H., Salas, J. D., Sanders, T. G., and Smith, R. E., editors, *Modeling Hydrologic Processes*. Water Resources Publications, Littleton, Colorado.

Salas, J. D., Tabios III, G. Q., and Bartolini, P. (1985). Approaches to multivariate modeling of water resources time series. *Water Resources Bulletin*, 21(4).

Shea, B. L. (1989). AS242 the exact likelihood of a vector autoregressive moving average model. *Applied Statistics*, 38(1):161-184.

Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G. (1986). *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium* held at Colorado State University, Fort Collins, Colorado, July 15-17, 1985, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

Stedinger, J. R. (1981). Estimating correlations in multivariate streamflow models. *Water Resources Research*, 17(1):200-208.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association*, 76:802-816.

Tiao, G. C. and Tsay, R. S. (1983a). Multiple time series modeling and extended cross-correlations. *Journal of Business and Economic Statistics*, 1:43-56.

Tiao, G. C. and Tsay, R. S. (1983b). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Annals of Statistics*, 11:856-871.

Tjostheim, D. and Paulsen, J. (1982). Empirical identification of multiple time series. *Journal of Time Series Analysis*, 3(4):265-282.

Tsay, R. S. and Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association*, 79:84-96.

Wilson, G. T. (1973). The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society*, Series B, 35(1):76-85.

Young, G. K. and Pisano, W. C. (1968). Operational hydrology using residuals. *Journal of the Hydraulics Division*, ASCE, 94(HY4):909-923.

NONGAUSSIAN MODELS

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26:211-252.

Brillinger, D. R. (1985). Fourier inference: Some methods for the analysis of array and nonGaussian series data. *Water Resources Bulletin*, 21(5):743-756.

Lewis, P. A. W. (1985). Some simple models for continuous variate time series. *Water Resources Bulletin*, 21(4):635-644.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin*, 21(4):645-650.

Li, W. K. and McLeod, A. I. (1988). ARMA modelling with nonGaussian innovations. *Journal of Time Series Analysis*, 9(2):155-168.

NONLINEAR MODELS

Gallant, A. R. (1987) *Nonlinear Statistical Models*. Wiley, New York.

Li, W. K. (1992). On the asymptotic standard errors residual autocorrelations in nonlinear time series modelling. *Biometrika*, 79(2):435-437.

Li, W. K. (1993). A simple one degree of freedom test for non-linear time series model discrimination. *Statistica Sinica*, 3(1):245-254.

Ozaki, T. (1985). Statistical identification of storage models with application to stochastic hydrology. *Water Resources Bulletin*, 21(4).

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag, New York.

Tong, H. (1990). *Nonlinear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

Tong, H., Thanoon, B., and Gudmundsson, G. (1985). Threshold time series modelling of two Icelandic riverflow systems. *Water Resources Bulletin*, 21(4):651-661.

PATTERN RECOGNITION

MacInnes, C. D. and Unny, T. E. (1986). Multivariate hydrologic time series and pattern recognition. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, held at Colorado State University, Fort Collins, Colorado, July 15-17, 1985, Engineering Research Center, Colorado State University, pages 228-242.

Panu, U. S. and Unny, T. E. (1980). Stochastic synthesis of hydrologic data based on concepts of pattern recognition, Parts I, II and III. *Journal of Hydrology*, 46:5-34, 197-217, and 219-237.

Unny, T. E., Panu, U. S., MacInnes, C. D., and Wong, A. K. C. (1981). Pattern analysis and synthesis of time-dependent hydrologic data. *Advances in Hydroscience*, 12:195-295.

PERIODIC MODELS

Bartolini, P. and Salas, J. D. (1986). Properties of multivariate periodic ARMA(1,1) processes. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, Fort Collins, Colorado, July 15-17, 1985, Engineering Research Center, Colorado State University, pages 264-282.

Bartolini, P., Salas, J. D., and Obeysekera, J. T. B. (1988). Multivariate periodic ARMA(1,1) processes. *Water Resources Research*, 24(8):1237-1246.

Haltiner, J. P. and Salas, J. D. (1988). Development and testing of a multivariate, seasonal ARMA(1,1) model. *Journal of Hydrology*, 104:247-272.

Hipel, K. W. and McLeod, A. I. (1981). Box-Jenkins modelling in the geophysical sciences. In Craig, R. G. and Labovitz, M. L., editors, *Future Trends in Geomathematics*, pages 65-86. Prion, United Kingdom.

Salas, J. D. and Abdelmohsen, M. W. (1993). Initialization for generating single-site and multisite low-order periodic autoregressive and moving average processes. *Water Resources Research*, 29(6):1771-1776.

Salas, J. D. and Pegram, G. G. S. (1978). A seasonal multivariate multilag autoregressive model in hydrology. In Morel-Seytoux, H., Salas, J. D., Sanders, T. G., and Smith, R. E., editors, *Modelling Hydrologic Processes*. Water Resources Publications, Littleton, Colorado.

Thomas, H. A. and Fiering, M. B. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Maass, A., Hufshmidt, M. M., Dorfman, R., Thomas Jr., H. A., Marglin, S. A., and Fair, M. G., editors, *Design of Water Resources Systems*, Harvard University Press, Cambridge, Massachusetts, pages 459-493.

Thompstone, R. M., Hipel, K. W., and McLeod, A. I. (1985a). Forecasting quarter-monthly riverflow. *Water Resources Bulletin*, 21(5):731-741.

Thompstone, R. M., Hipel, K. W., and McLeod, A. I. (1985b). Grouping of periodic autoregressive models. In Anderson, O. D., Ord, J. K., and Robinson, E. A., editors, *Time Series Analysis: Theory and Practice* 6. North-Holland, Amsterdam, pages 35-49.

Ula, T. A. (1990). Periodic covariance stationarity of multivariate periodic autoregressive moving average processes. *Water Resources Research*, 26(5):855-861.

Vecchia, A. V. (1985a). Periodic autoregressive-moving average (PARMA) modeling with applications to water resources. *Water Resources Bulletin*, 21(5):721-730.

Vecchia, A. V. (1985b). Maximum likelihood estimation for periodic autoregressive-moving average models. *Technometrics*, 27:375-384.

Vecchia, A. V., Obeysekera, J. T., Salas, J. D., and Boes, D. C. (1983). Aggregation and estimation for low-order periodic ARMA models. *Water Resources Research*, 19(5):1297-1306.

PHYSICAL BASIS OF MODELS

Klemes, V. (1978). Physically based stochastic hydrologic analysis. In Chow, V. T., editor, *Advances in Hydroscience*, 11:285-356.

Salas, J. D. and Smith, R. A. (1981). Physical basis of stochastic models of annual flows. *Water Resources Research*, 17(2):428-430.

Yevjevich, V. and Harmancioglu, N. B. (1985). Past and future of analysis of water resources time series. *Water Resources Bulletin*, 21(4):625-633.

SIMULATION

Davis, D. R. (1977). Comment on Bayesian generation of synthetic streamflows by G. J. Vicens, I. Rodriguez-Iturbe and J. C. Schaake Jr. *Water Resources Research*, 13(5):853-854.

Feldman, A. D. (1981). HEC models for water resources system simulation: Theory and experience. *Advances in Hydroscience*, 12:297-423.

Finzi, G., Todini, E., and Wallis, J. R. (1975). Comment upon multivariate synthetic hydrology. *Water Resources Research*, 11(6):844-850.

Kelman, J. (1980). A stochastic model for daily streamflow. *Journal of Hydrology*, 47:235-249.

Kelman, J., Damazio, J. M., and Costa, J. P. (1986). A multivariate synthetic daily streamflow generator. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, July 15-17, 1985, Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 121-140.

Kottegoda, N. T. and Yevjevich, V. (1977). Preservation of correlation in generated hydrologic samples through two-station models. *Journal of Hydrology*, 33:99-121.

McLeod, A. I. and Hipel, K. W. (1978c). Simulation procedures for Box-Jenkins models. *Water Resources Research*, 14(5):969-975.

Mejia, J. M., Rodriquez-Irturbe, I., and Cordova, J. R. (1974). Multivariate generation of mixtures of normal and log normal variables. *Water Resources Research*, 10(4):691-693.

Nasseri, I. (1986). Multistation stochastic model of hourly rainfall. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, July 15-17, 1985, Fort Collins, Colorado, Engineering Research Center, Colorado State University, pages 531-541.

Pegram, G. G. S. and James, W. (1972). Multilag multivariate autoregressive model for the generation of operational hydrology. *Water Resources Research*, 8(4):1074-1076.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature and solar radiation data. *Water Resources Research*, 17(1):182-190.

Srikanthan, R. (1986). Stochastic simulation of daily climatic data. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, July 15-17, 1985, Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 542-554.

Stedinger, J. R. and Taylor, M. R. (1982a). Synthetic streamflow generation, 1, Model verification and validation. *Water Resources Research*, 18(4):909-918.

Stedinger, J. R. and Taylor, M. R. (1982b). Synthetic streamflow generation, 2, Parameter uncertainty. *Water Resources Research*, 18(4):919-924.

Valdes, J. B., Rodriguez-Iturbe, I., and Vicens, G. J. (1977). Bayesian generation of synthetic streamflows, 2, Multivariate case. *Water Resources Research*, 13(2):291-295.

Venugopal, K., Ranganathan, T., Babu Rao, T., Raghavendran, R., and Sakthivadivel, R. (1986). Stochastic multisite modelling of Narmada River system in India. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Symposium*, July 15-17, 1985, Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 664-677.

Vicens, G. J., Rodriguez-Iturbe, I., and Schaake Jr., J. C. (1975). Bayesian generation of synthetic stream-flows. *Water Resources Research*, 11(6):827-838.

SPECTRUM

Canfield, R. V. and Bowles, D. S. (1986). Multivariate series generation from the spectral density function. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985, Engineering Research Center, University of Colorado, Fort Collins, Colorado, pages 572-581.

Ghani, A. A. A. and Metcalfe, A. V. (1986). Spectral predictions of flood risk. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985, Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 744-754.

Jenkins, G. M. and Watts, D. G. (1968). Spectral Analysis and its Applications. Holden-Day, San Francisco.

STARMA MODELS

Adamowski, K., Mohamed, F. B., Dalezios, N. R. and Birta, L. G. (1986). Space-time ARIMA modelling of precipitation time series. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985. Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 217-227.

Deutsch, S. J. and Pfeifer, P. E. (1981). Space-time ARMA modelling with contemporaneously correlated innovations. *Technometrics,* 23(4):401-409.

Deutsch, S. J. and Ramos, J. A. (1984). Space time evaluation of reservoir regulation policies. Technical report, School of Civil Engineering, Georgia Institute of Technology.

Deutsch, S. J. and Ramos, J. A. (1986). Space-time modeling of vector hydrologic sequences. *Water Resources Bulletin,* 22(6):967-981.

Pfeifer, P. E. and Deutsch, S. J. (1980). Identification and interpretation of first-order space-time ARMA models. *Technometrics,* 22:397-408.

TFN MODELS

Delleur, J. W. (1986). Recursive parameter identification for flashflood forecasting. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985. Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 154-177.

Nicklin, M. E. (1986). Bivariate transfer function modeling of irrigation return flows. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985. Engineering Research Center, Colorado State University, Fort Collins, Colorado, pages 620-637.

CHAPTER 21

CONTEMPORANEOUS AUTOREGRESSIVE-MOVING AVERAGE MODELS

21.1 INTRODUCTION

The *contemporaneous ARMA, or CARMA, family of models* is designed for modelling two or more time series that are statistically related to one another only at the same time, or simultaneously. For example, two riverflow series that are measured within the same climatic zone but not at locations where one station is upstream from the other, may be only correlated simultaneously with one another. As demonstrated by the applications given in Section 21.5 of this chapter, a CARMA model is the most appropriate type of multivariate model to describe this situation mathematically.

Because of the usefulness of CARMA modelling in water resources, this chapter is devoted entirely to presenting this interesting and simple model. As explained in Section 20.2.2 of the previous chapter, CARMA models actually form a special type of *general multivariate ARMA models*. Besides possessing far fewer parameters than the general multivariate ARMA models described in detail in Chapter 20, CARMA models can be conveniently fitted to multiple time series using well developed model construction techniques.

Another useful subset of models from the general multivariate ARMA family is the group of *TFN models* which includes the closely related *intervention models*. TFN models can be employed when a single output series is dependent upon one or more input series plus a noise component. If a single output series is affected by one or more external interventions and perhaps also some input series, an intervention model can parsimoniously describe this situation. Along with many interesting applications, TFN models are presented in Part VII while intervention models are discussed in detail in Part VIII and Section 22.4.

Descriptions of the historical development of multivariate models in water resources are presented in Sections 20.4 and 20.5 as well as in the papers by Salas et al. (1985) and Hipel (1986). In addition to other types of multivariate models, many references are listed at the end of Chapter 20 for previous research in CARMA modelling. Much of the material presented in this chapter is drawn from research completed by Camacho et al. (1985, 1986, 1987a,b,c) and Camacho (1984).

In the next section, two alternative approaches to deriving the equations for CARMA models are presented. Following this, a comprehensive set of *model construction tools* are described in Section 21.3. To avoid introducing bias into synthetic sequences, a correct method for *generating simulated data* from a CARMA model is presented in Section 21.4. The *practical applications* in Section 21.5 demonstrate how convenient and simple it is to use the building methods for properly describing both water quantity and quality time series.

21.2 DERIVING CARMA MODELS

21.2.1 Introduction

CARMA models can be defined using two distinct viewpoints. Firstly, as noted in Section 20.2.2, the CARMA group of models can be thought of as being a subset of the general multivariate ARMA family of models. Instead of going from a more general class of models to a more specific subset of models, the second approach for defining the CARMA group of models goes in the reverse direction. In particular, a CARMA model can be considered as a collection of, say, k univariate ARMA models with contemporaneously correlated innovations. This second interpretation is particularly useful for the development of model construction tools, especially computationally efficient estimation algorithms. The subset and concatenation definitions of the CARMA group of models are now presented.

21.2.2 Subset Definition

The mathematical definition for the general multivariate ARMA family of models is given in [20.2.1] and [20.2.2]. By constraining the AR and MA parameter matrices to be diagonal matrices, the CARMA subset of models is defined. More specifically, following the notation of Section 20.2.2, let k time series at time t be represented by the vector $\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$ where the vector of the theoretical means for $\mathbf{Z}_t$ is given by $\mu = (\mu_1, \mu_2, \ldots, \mu_k)^T$. Assuming that the orders of the AR and MA components are p and q, respectively, the *CARMA(p,q) model* can be written as

$$(\mathbf{Z}_t - \mu) - \begin{bmatrix} \phi_{111} & & & & \\ & \phi_{221} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \phi_{kk1} \end{bmatrix} (\mathbf{Z}_{t-1} - \mu) - \begin{bmatrix} \phi_{112} & & & & \\ & \phi_{222} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \phi_{kk2} \end{bmatrix} (\mathbf{Z}_{t-2} - \mu)$$

$$- \cdots - \begin{bmatrix} \phi_{11p} & & & & \\ & \phi_{22p} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \phi_{kkp} \end{bmatrix} (\mathbf{Z}_{t-p} - \mu) \qquad [21.2.1]$$

$$= \mathbf{a}_t - \begin{bmatrix} \theta_{111} & & & \\ & \theta_{221} & & \\ & & \ddots & \\ & & & \theta_{kk2} \end{bmatrix} \mathbf{a}_{t-1} - \begin{bmatrix} \theta_{112} & & & \\ & \theta_{222} & & \\ & & \ddots & \\ & & & \theta_{kk2} \end{bmatrix} \mathbf{a}_{t-2} - \cdots$$

$$- \begin{bmatrix} \theta_{11q} & & & \\ & \theta_{22q} & & \\ & & \ddots & \\ & & & \theta_{kkq} \end{bmatrix} \mathbf{a}_{t-q}$$

where

$$\Phi_i = \begin{bmatrix} \phi_{11i} & & & \\ & \phi_{22i} & & \\ & & \ddots & \\ & & & \phi_{kki} \end{bmatrix}$$

is the AR parameter matrix for $i = 1,2,\ldots,p$, having zero entries for all the off diagonal elements;

$$\Theta_i = \begin{bmatrix} \theta_{11i} & & & \\ & \theta_{22i} & & \\ & & \ddots & \\ & & & \theta_{kki} \end{bmatrix}$$

is the MA parameter matrix for $i = 1,2,\ldots,q$, possessing zero values for all the non-diagonal elements; and

$$\mathbf{a}_t = (a_{t1}, a_{t2}, \ldots, a_{tk})^T$$

is the k dimensional vector of innovations for $\mathbf{Z}_t$ at time t. Notice that the model in [21.2.1] has the same form as the general multivariate ARMA model in [20.2.1] and [20.2.2], except for the fact that the AR and MA parameter matrices are diagonal.

After performing the matrix multiplications in [21.2.1], one obtains a set of k simultaneous difference equations. In particular, the ith difference equation for the variable Z_{ti} is

$$Z_{ti} - \mu_i - \phi_{ii1}(Z_{t-1,i} - \mu_i) - \phi_{ii2}(Z_{t-2,i} - \mu_i) - \cdots - \phi_{iip}(Z_{t-p,i} - \mu_i)$$
$$= a_{ti} - \theta_{ii1}a_{t-1,i} - \theta_{ii2}a_{t-2,i} - \cdots - \theta_{iiq}a_{t-q,i}\,, \quad \text{for } i = 1,2,\ldots,k \qquad [21.2.2]$$

Notice from [21.2.2] that only the ith variable and ith innovation series appear in the equation. The simultaneous correlation among the k variables is incorporated into the CARMA model by allowing the innovations to be contemporaneously correlated. More precisely, the vector of innovations given by $\mathbf{a}_t$ are assumed to be IID vector random variables with a mean of zero and variance covariance matrix given by $\Delta = E[\mathbf{a}_t \cdot \mathbf{a}_t^T]$. For practical applications, the normality assumption is invoked and $\mathbf{a}_t \sim NID(\mathbf{0},\Delta)$.

The model in [21.2.2] can be more compactly written as

$$\phi_i(B)(Z_{ti} - \mu_i) = \theta_i(B)a_{ti}\,, \quad i = 1,2,\ldots,k \qquad [21.2.3]$$

where

$$\phi_i(B) = 1 - \phi_{ii1}B - \phi_{ii2}B^2 - \cdots - \phi_{iip}B^p$$

is the ith AR operator of order p and

$$\theta_i(B) = 1 - \theta_{ii1}B - \theta_{ii2}B^2 - \cdots - \theta_{iiq}B^q$$

is the ith MA operator of order q. For the CARMA model to be stationary and invertible, the zeroes of the characteristic equations $\phi_i(B) = 0$ and $\theta_i(B) = 0$, respectively, must lie outside the unit circle.

Example: Consider a bivariate CARMA(1,1) model for connecting the two variables contained in the vector

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2})^T$$

having theoretical means given by

$$\mu = (\mu_1, \mu_2)^T$$

From [21.2.1], the bivariate CARMA(1,1) model is written as

$$\begin{bmatrix} Z_{t1} - \mu_1 \\ Z_{t2} - \mu_2 \end{bmatrix} - \begin{bmatrix} \phi_{111} & 0 \\ 0 & \phi_{221} \end{bmatrix} \begin{bmatrix} Z_{t-1,1} - \mu_1 \\ Z_{t-1,2} - \mu_2 \end{bmatrix} = \begin{bmatrix} a_{t1} \\ a_{t2} \end{bmatrix} - \begin{bmatrix} \theta_{111} & 0 \\ 0 & \theta_{221} \end{bmatrix} \begin{bmatrix} a_{t-1,1} \\ a_{t-1,2} \end{bmatrix} \qquad [21.2.4]$$

After matrix multiplication, the two component equations of the bivariate model are

$$Z_{t1} - \mu_1 - \phi_{111}(Z_{t-1,1} - \mu_1) = a_{t1} - \theta_{111}a_{t-1,1}$$
$$Z_{t2} - \mu_2 - \phi_{221}(Z_{t-1,2} - \mu_2) = a_{t2} - \theta_{221}a_{t-1,2} \qquad [21.2.5]$$

The vector of innovations for the bivariate model is

$$\mathbf{a}_t = (a_{t1}, a_{t2})^T$$

where the variance covariance matrix for $\mathbf{a}_t$ is

$$\Delta = E[\mathbf{a}_t \cdot \mathbf{a}_t^T] = E\left[\begin{pmatrix} a_{t1} \\ a_{t2} \end{pmatrix}(a_{t1}\ a_{t2})\right]$$

$$= \begin{bmatrix} E[a_{t1}^2] & E[a_{t1}a_{t2}] \\ E[a_{t2}a_{t1}] & E[a_{t2}^2] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Because $\sigma_{21} = \sigma_{12}$, then Δ is a symmetric matrix. Under the normality assumption $\mathbf{a}_t \sim NID(\mathbf{0},\Delta)$.

To satisfy the stationarity assumption, the roots of

$$\phi_1(B) = 1 - \phi_{111}B = 0$$

and

$$\phi_2(B) = 1 - \phi_{221}B = 0$$

must lie outside the unit circle. Consequently, $|\phi_{111}| < 1$ and $|\phi_{221}| < 1$. For invertibility, the roots of

$$\theta_1(B) = 1 - \theta_{111}B = 0$$

and

$$\theta_2(B) = 1 - \theta_{221}B = 0$$

must lie outside the unit circle. Hence, $|\theta_{111}| < 1$ and $|\theta_{221}| < 1$ for satisfying the invertibility condition.

21.2.3 Concatenation Definition

The clue to discovering the second approach to defining a CARMA model is given by the form of the component equation in [21.2.2] and [21.2.3]. Notice that the model in [21.2.3] for the ith variable is in fact an ARMA model and is identical to the ARMA model defined in [3.4.3] and [3.4.4]. Accordingly, one can consider the CARMA model to consist of a concatenation of k ARMA models where there is a separate ARMA model to describe each of the k series. In general, the orders of the AR and MA operators may vary across the k models. Therefore, the ARMA model for Z_{ti} can be written more precisely as

$$\phi_i(B)(Z_{ti} - \mu_i) = \theta_i(B)a_{ti}\ ,\ i = 1,2,\ldots,k \qquad [21.2.6]$$

where

$$\phi_i(B) = 1 - \phi_{ii1}B - \phi_{ii2}B^2 - \cdots - \phi_{iip_i}B^{p_i}$$

is the ith AR operator of order p_i and

$$\theta_i(B) = 1 - \theta_{ii1}B - \theta_{ii2}B^2 - \cdots - \theta_{iiq_i}B^{q_i}$$

is the ith MA operator of order q_i. The chain that links the k ARMA models together in terms of contemporaneous correlation is the variance covariance matrix, Δ, for the innovations $\mathbf{a}_t = (a_{t1}, a_{t2}, \ldots, a_{tk})^T$ where $\mathbf{a}_t \sim NID(\mathbf{0}, \Delta)$. When using the notation CARMA(p,q) to stand for the overall model, one sets $p = \max(p_1, p_2, \ldots, p_k)$ and $q = \max(q_1, q_2, \ldots, q_k)$. By constraining appropriate parameters to be zero in the subset definition of the CARMA(p,q) model in [21.2.1], one can also allow the orders of the AR and MA operators to vary when using this equivalent definition.

In summary, from [21.2.6], the CARMA model can be thought of as a set of k univariate ARMA models for which the innovations are contemporaneously correlated. This contemporaneous correlation is modelled using the variance covariance matrix, Δ, which has a typical entry denoted by σ_{ij}. For the situation where none of the series are contemporaneously correlated with each other, $\sigma_{ij} = 0$ for $i \neq j$ and the multivariate CARMA model collapses into a collection of k independent univariate ARMA models. Consequently, one can interpret the CARMA model as a natural extension of the univariate ARMA model. Alternatively, under the subset definition in Section 21.2.2, the CARMA model can be considered to be a special case of a more general family of models.

21.3 CONSTRUCTING CARMA MODELS

21.3.1 Introduction

Because flexible and simple model construction procedures are now available for fitting CARMA models to a data set, it is currently possible for practitioners to conveniently employ these models in practical applications. Some of the techniques used at the three stages of model construction have naturally evolved from the concatenation interpretation of the CARMA model presented in [21.2.6]. Consequently, construction methods used for fitting univariate ARMA models have been cleverly extended for use with CARMA models for which there are contemporaneous correlations among the innovation series. Specific details and a comprehensive list of references regarding the available procedures for use in model fitting can be found in papers by authors such as Camacho et al. (1985, 1986, 1987a,b,c), Salas et al. (1985), Hipel (1986) and Jenkins and Alavi (1981). In this section, some of the most useful model construction tools are described.

21.3.2 Identification

A sound *physical understanding* of a given problem in conjunction with a thorough appreciation of the capabilities of the various types of multivariate ARMA models, are of utmost importance in model identification. For instance, when riverflows from different river basins are controlled by the same general climatic conditions within an overall region, a CARMA model may be appropriate to use with this multisite data. The *residual CCF* is a very useful statistical tool for ascertaining statistically whether or not a CARMA model is needed to fit to two or more time series and also to decide upon the orders of the AR and MA parameters. In Section 16.2.2, the theoretical and sample residual CCF functions are defined in [16.2.5] and [16.2.6], respectively, and it is explained how the residual CCF can be used to determine the type of causality existing between two series and thereby confirm in a statistical manner what one may suspect a priori from a physical understanding of the problem. A summary of the use of the residual CCF in causality studies is given in Table 16.2.1 and also in Section 20.3.2 under the heading Causality.

Suppose that one has a set of k time series given by

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$$

where each series has n equally spaced observations that are available at the same time as the other series. For the ith time series, the data set is given as $[Z_{1i}, Z_{2i}, \ldots, Z_{ni}]$. Using the sample residual CCF for model identification involves the following two steps.

Step 1 - Fitting Univariate ARMA Models: Using the ARMA model construction procedures of Part III, the most appropriate ARMA model is fitted separately to each of the data sets $\left\{Z_{1i}, Z_{2i}, \ldots, Z_{ni}\right\}$, $i = 1,2, \ldots, k$. This step produces a residual series

$$\left\{\bar{a}_{ti}\right\} = \left\{\bar{a}_{1i}, \bar{a}_{2i}, \ldots, \bar{a}_{ni}\right\}$$

for each series $i = 1,2, \ldots, k$, for the univariate ARMA model in [21.2.6] or [3.4.3]. Obtaining the residuals of an ARMA model fitted to a given series is referred as prewhitening in Sections 16.2.2 and 20.3.2. Besides the residuals, a vector of parameter estimates given by

$$\bar{\beta}_i = (\bar{\phi}_{ii1}, \bar{\phi}_{ii2}, \ldots, \bar{\phi}_{iip_i}, \bar{\theta}_{ii1}, \bar{\theta}_{ii2}, \ldots, \bar{\theta}_{iiq_i})^T$$

is found for each of the series $i = 1,2, \ldots, k$. The bar above a variable or parameter means that the variable or parameter has been estimated using an efficient univariate estimation procedure from Section 6.2.3 or Appendix A6.1.

Step 2 - Analysis of the Residual CCF: As explained in Section 16.2.2, to determine statistically the type of causality existing between two series Z_{ti} and Z_{tj}, one examines the *residual CCF* which is calculated for the residuals series $[\bar{a}_{ti}]$ and $[\bar{a}_{tj}]$. Following [16.2.6], the residual CCF is determined for lag l as

$$\bar{r}_{ij}(l) = \bar{c}_{ij}(l) \,/\, [\bar{c}_{ii}(0)\bar{c}_{jj}(0)]^{1/2}$$

where

$$\bar{c}_{ij}(l) = \begin{bmatrix} n^{-1}\sum_{t=1}^{n-l} \bar{a}_{ti}\bar{a}_{t+l,j}\,, & \text{for } l \geq 0 \\ n^{-1}\sum_{t=1-l}^{n} \bar{a}_{ti}\bar{a}_{t+l,j}\,, & \text{for } l < 0 \end{bmatrix} \qquad [21.3.1]$$

is the estimated cross covariance function at lag l between the two residual series, and $\bar{c}_{ii}(0)$ and $\bar{c}_{jj}(0)$ are the estimated variances of the ith and jth residual series, respectively.

The residual CCF can be calculated for negative, zero and positive lags for all possible pairs of series. If a CARMA model is adequate for modelling the data, only the residual CCF at lag zero should be significantly different from zero. If this is not the case, a more complicated model such as a TFN model (see Chapter 17) or a general multivariate ARMA model (see Chapter 20) may be needed. Under the hypothesis that the CARMA model is adequate for describing the data, the quantities $\pm 2/n^{1/2}$ can be considered as approximate 95% confidence limits to decide whether a value of the residual CCF is significant or not. The test for the significance of the cross correlations can be easily performed by plotting the residual CCF

$$\bar{r}_{ij}(l),\ l = 0, \pm 1, \pm 2, \ldots, \pm m$$

where $m < n/4$ together with the 95% confidence limits for each distinct pair of residual series. If a CARMA model is appropriate for modelling the series, only the residual CCF at lag zero will be significantly different from zero for all pairs of series.

An alternative to plotting the residual CCF's is to summarize the significance of each value of the residual CCF in the *residual CCF matrix* denoted by $\overline{\mathbf{R}}(l) = [\bar{r}_{ij}(l)]$. Because there are k series, the dimension of $\overline{\mathbf{R}}(l)$ is $k \times k$ where the (i,j) entry gives the result for the ith and jth series. Also, since $\bar{r}_{ij}(l) = -\bar{r}_{ji}(l)$, one only has to determine the residual CCF matrices $\overline{\mathbf{R}}(l)$ for zero and positive lags so that $l = 0, 1, 2, \ldots, m$. For convenience in detecting significant values in $\overline{\mathbf{R}}(l)$, each $\bar{r}_{ij}(l)$ entry can be replaced by a "+" to indicate a value greater than $2n^{-1/2}$ or or by a "-" to point out a value smaller than $-2n^{-1/2}$ or by a "." to indicate a value falling between $-2n^{-1/2}$ and $2n^{1/2}$. Thus, "+", "-" and "." stand for values significantly greater, significantly less and not significantly different from zero, respectively. If the approximate 95% confidence interval given by $(-2n^{1/2}, 2n^{1/2})$ is not considered to be accurate enough, exact confidence intervals could be calculated (Li and McLeod, 1981), although this is not usually necessary.

In summary, from Step 1, one knows the number of AR and MA parameters required to model each series and one has univariate estimates of these parameters. If the residual CCF calculated in Step 2 for each pair of series is only significantly different from zero at lag zero, then a CARMA model is the most appropriate type of multivariate model to fit to the k series. An advantage of using the residual CCF as an identification technique is that it may indicate the direction of departure from the CARMA model, if this model is not adequate to fit the data. For example, if the residual CCF were significantly different from zero for lag zero and also a few positive lags but not significant for any negative lags, this may indicate that a TFN model is required (see Part VII). Although one could also use the model identification techniques described in Appendix A20.1 of the previous chapter, these techniques are not as convenient to use as the residual CCF, especially when one suspects from a physical viewpoint that a CARMA model is needed.

21.3.3 Estimation

After a tentative model has been identified, the next step is to estimate the parameters of the model. General multivariate ARMA estimation procedures based upon maximum likelihood, such as the methods of Hillmer and Tiao (1979) and Nicholls and Hall (1979) referred to in Section 20.3.2, could be employed to estimate the parameters of the CARMA model. It should be pointed out, however, that these algorithms are not computationally efficient for the estimation of the parameters of the CARMA model, and efficient algorithms can be readily obtained, as explained below. On the other hand, the univariate estimates $\bar{\beta}_i$ obtained for each of the series Z_{ti}, $i = 1,2, \ldots, k$ in Step 1 in Section 21.3.2, do not provide statistically efficient estimators of the parameters in the overall CARMA model. This is because the variance of the estimated AR and MA parameters contained in $\bar{\beta}_i$ for the ith time series may be quite high. Camacho (1984) and Camacho et al. (1987a,b) show theoretically that the variances of the univariate estimators $\bar{\beta}_i$, $i = 1,2, \ldots, k$, are greater than the variances of the estimators obtained using the joint multivariate estimation algorithm described below. In some cases, the univariate estimators are much less efficient than the joint multivariate estimators.

To overcome the aforementioned inefficiencies of the estimation techniques, Camacho et al. (1987a,b) developed an algorithm to obtain efficient MLE's of the model parameters. As is also assumed at the identification stage, let $\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$ for $t = 1,2, \ldots, n$ be a sample of n consecutive observations for the k time series Z_{ti}, $i = 1,2, \ldots, k$. Hence, for the ith time series the set of observations is given as

$$\{Z_{ti}\} = \{Z_{1i}, Z_{2i}, \ldots, Z_{ni}\}$$

Let the parameters of the CARMA model for the ith series be contained in the vector

$$\beta_i = (\phi_{ii1}, \phi_{ii2}, \ldots, \phi_{iip_i}, \theta_{ii1}, \theta_{ii2}, \ldots, \theta_{iiq_i})^T$$

Consequently, the vector of parameters for the complete CARMA model is written as

$$\beta = (\beta_1, \beta_2, \ldots, \beta_k)^T$$

The CARMA estimation algorithm consists of the following steps:

1. For each series Z_{ti}, $i = 1,2, \ldots, k$, obtain univariate estimates of the ARMA model parameters using univariate ARMA estimation techniques such as those by Newbold (1974), Ansley (1979), Ljung and Box (1979), and McLeod (1977) referred to in Section 6.2.3. The ARMA estimator of McLeod is described in Appendix A6.1. In Step 1 of the identification stage of the previous section, the univariate estimates $\bar{\beta}_i$, $i = 1,2, \ldots, k$, are already found in order to produce the prewhitened series for each fitted ARMA model. Recall that a bar written above a vector indicates that an efficient univariate estimator has been employed to obtain estimates of the parameters contained in the vector.

2. Calculate

$$\beta^* = \bar{\beta} - V(\bar{\beta})(\partial S/\partial\beta)\Big|_{\beta=\bar{\beta}} \qquad [21.3.2]$$

where $\bar{\beta} = (\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k)^T$ is the vector of univariate estimates for which $\bar{\beta}_i$ is the vector of univariate estimates of the ARMA model for the ith series;

$$V(\bar{\beta})^{-1} = plim(\partial^2 S/\partial\beta\partial\beta^T)\Big|_{\beta=\bar{\beta}}$$

is the inverse of the variance-covariance matrix for the parameters contained in $\bar{\beta}$, which is the information matrix;

$$(\partial S/\partial \beta)\Big|_{\beta=\bar{\beta}}$$

denotes the vector of partial derivates of the sum of squares function, S, with respect to the CARMA parameter β, evaluated at the point $\beta=\bar{\beta}$. The sum of squares function S, is defined as

$$S = \sum_{t=1}^{n} \mathbf{a}_t^T \Delta^{-1} \mathbf{a}_t / 2n$$

where

$$\mathbf{a}_t = (a_{t1}, a_{t2}, \ldots, a_{tk})^T,$$

$$\Delta = (\sigma_{ij})$$

is the variance covariance matrix of $\mathbf{a}_t$, and a_{ti} is defined in

$$a_{ti} = Z_{ti} - \phi_{ii1} Z_{t-1,i} - \phi_{ii2} Z_{t-2,i} - \cdots - \phi_{iip_i} Z_{t-p_i,i}$$
$$+ \theta_{ii1} a_{t-1,i} + \theta_{ii2} a_{t-2,i} + \cdots + \theta_{iiq_i} a_{t-q_i,i}, \quad t \geq p_i$$

Initial values for a_{ti} can be calculated using the algorithm given by McLeod and Sales (1983) or can be set equal to zero. To calculate the information matrix or, equivalently, $\mathrm{V}(\beta)^{-1}$ which is the inverse of the variance-covariance matrix of β, the algorithm given by Ansley (1980) and Kohn and Ansley (1982) can be employed. Camacho (1984) proves that β^* is asymptotically efficient. Using $\bar{\beta}$ as the initial point, the estimation procedure corresponds to one iteration of the Gauss Newton optimization scheme. To obtain the maximum likelihood estimator, $\hat{\beta}$, for the complete CARMA model, iterations can be continued until convergence is reached.

Camacho et al. (1985, 1987a) extend their estimation algorithm for CARMA models to include the situation where the multiple time series have unequal sample sizes. In this way, the modeller can take full advantage of all the available data and none of the observations in any of the series have to be omitted from the analysis. This estimation algorithm is outlined in Appendix A21.1.

Camacho et al. (1986) consider the effect on the estimation of the parameters when a bivariate series $\mathbf{Z}_t = (Z_{t1}, Z_{t2})^T$ is incorrectly modelled as a general multivariate AR(1) model using [20.2.1] when a CARMA(1,0) model from [21.2.1] would suffice. As pointed out in Section 20.4, the general multivariate AR(1) model has been proposed for utilization in hydrology. Using simulation studies, they show that the loss in efficiency of the parameter estimates obtained using the full multivariate model can be very substantial and in many cases can be well over 50%.

21.3.4 Diagnostic Checks

After obtaining efficient estimates for the model parameters, possible inadequacies in the fitted model can be found and subsequently corrected by examining the statistical properties of the residuals. As explained in Section 20.3.2, a range of tests are available for ascertaining whether or not the residuals are *white* (Li and McLeod, 1981), *homoscedastic* (see Section 7.5.2) and *normally distributed* (Royston, 1983).

For detecting misspecifications in the model, the *residual CCF* is both informative and sensitive. In addition to the joint estimates for the model parameters, one can obtain the model residuals $\hat{a}_{ti}$, $i = 1,2,\ldots,k$, using the efficient estimation procedure of Section 21.3.3. To calculate the residual CCF, $\hat{r}_{ij}(l)$, at lag l between two residual series, one simply replaces $\bar{a}_{ti}$ and $\bar{a}_{tj}$ by $\hat{a}_{ti}$ and $\hat{a}_{tj}$, respectively, in [21.3.1]. Each entry in the *residual CCF matrix*, $\hat{\mathbf{R}}(l)$, should not be significantly different from zero for $l > 1$. As is done at the identification stage, it is convenient to use the symbols "+", "-", and "." in $\hat{\mathbf{R}}(l)$ to indicate entries that are significantly larger, significantly smaller, and not significantly different from zero, respectively.

Based upon the work of Li and McLeod (1981), Camacho et al. (1985) suggest a *modified Portmanteau test* statistic to test for the independence of the residuals. As in Section 21.3.3, let $\Delta = E[\mathbf{a}_t \cdot \mathbf{a}_t^T]$ be the variance-covariance matrix of $\mathbf{a}_t = (a_{t1}, a_{t2}, \ldots, a_{tk})^T$ and let

$$\hat{r}(l) = (\hat{r}_{11}(l), \hat{r}_{21}(l), \ldots, \hat{r}_{k1}(l), \hat{r}_{12}(l), \hat{r}_{22}(l), \ldots, \hat{r}_{k2}(l), \ldots, \hat{r}_{kk}(l))^T$$

The modified Portmanteau test statistic is then written as

$$Q_{L_m} = n \sum_{l=1}^{L} \hat{r}(l)^T (\hat{\Delta}^{-1} \otimes \hat{\Delta}^{-1}) \hat{r}(l) + k^2 L(L+1)/2 \qquad [21.3.3]$$

Under the assumption that the residuals are white noise, Q_{L_m} is approximately χ^2 distributed with $k^2L - k(p+q)$ degrees of freedom for large values of L and n.

If the residual CCF possesses significantly large values at lags other than zero, the CARMA model must be appropriately redesigned. Perhaps it may be only necessary to add additional AR and MA parameters to the CARMA model. If the CARMA class of models itself is not adequate, a more complex family of multivariate models, such as the TFN set of models, may have to be considered. When the residuals are not approximately normally distributed and/or homoscedastic, it may be required to transform one or more of the series using an appropriate transformation such as the Box-Cox transformation in [3.4.30]. Subsequent to this, the parameters of the CARMA can be estimated again using the algorithm in [21.3.3].

21.3.5 Seasonality

The CARMA(p,q) model presented in Sections 21.2.2 and 21.2.3 is defined for handling nonseasonal series and the model construction techniques of this section are explained for the nonseasonal case. When one wishes to fit a CARMA model to seasonal data, the two approaches described in more detail in Section 20.3.3 can be used. In particular, one can first *deseasonalize* each series using a technique from Section 13.2.2 and then fit a nonseasonal CARMA model to the deseasonalized data. An alternative approach is to employ a *periodic version of the CARMA model* to fit directly to the seasonal data.

As explained in Part VI for univariate models, usually deseasonalized (Chapter 13) or periodic (Chapter 14) models are the most appropriate types of seasonal models to fit to natural time series. This is because data within a given season for a natural time series are usually stationary across the years. However, when the data within seasons are nonstationary over the years, it may be appropriate to seasonally and perhaps also nonseasonally difference the data to remove the nonstationarity (see Chapter 12). For example, a seasonal economic time series may possess an upward trend which causes the overall level of the series to increase over the years. After appropriately differencing the series, a seasonal ARMA model can be fitted to the resulting stationary data in order to obtain the parameter estimates for the seasonal ARIMA model.

In a manner similar to that for the univariate seasonal ARIMA model of Chapter 13, a *CARIMA model* containing differencing operators can be easily defined. To accomplish this, one simply introduces seasonal and nonseasonal differencing operators along with seasonal AR and MA operators into [21.2.7]. A model containing differencing operators could also be defined for the general multivariate ARMA models of the previous chapter.

21.4 SIMULATING USING CARMA MODELS

21.4.1 Introduction

Comprehensive techniques for generating synthetic sequences using ARMA and ARIMA models were developed by McLeod and Hipel (1978b) and are presented in detail in Chapter 9. To avoid the introduction of systematic bias into the simulated series by employing fixed starting values, the simulation methods described in Sections 9.3 and 9.4 are designed such that random realizations of the underlying model are used for starting values. The simulation techniques developed for the univariate ARMA and ARIMA models can be extended for use with CARMA models.

Originally, McLeod (1979) suggested a simulation algorithm for use with CARMA models possessing no MA parameters while Camacho (1984) presented the algorithm for the general case. Simulation experiments which employ this new algorithm are given by McLeod (1979), Camacho (1984), and Camacho et al. (1985, 1986). A similar type of simulation algorithm can be developed for use with the general multivariate ARMA models of Chapter 20.

21.4.2 Simulation Algorithm

Overall Algorithm

Suppose that there are k time series and at time t the vector of time series is denoted by

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tk})^T$$

For the ith time series, let the order of the AR and MA parameters needed in [21.2.6] be p_i and q_i, respectively. Now, define

$$\mathbf{Z}_{p_i,i} = (Z_{1i}, Z_{2i}, \ldots, Z_{p_i,i})^T$$

and

$$\mathbf{a}_{q_i,i} = (a_{p_i-q_i+1,i}, a_{p_i-q_i+2,i}, \ldots, a_{p_i,i})^T$$

for series $i = 1,2, \ldots, k$. Then, the values contained in the vectors $\mathbf{Z}_{p_i,i}$ and $\mathbf{a}_{q_i,i}$ represent the starting values for the ith series where $i = 1,2, \ldots, k$.

Suppose that it is required to generate N synthetic observations for the CARMA model in [21.2.6]. Without loss of generality, it is assumed that the mean of each of the k series is zero. The following algorithm provided by Camacho (1984, pp. 57-68) is used to obtain simulated values $\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_N$ where $\mathbf{Z}_t=(Z_{t1},Z_{t2},\ldots,Z_{tk})^T$. Moreover, this algorithm is exact in the sense that it is not subject to inaccuracies associated with fixed initial values.

1. Determine the lower triangular matrix $\mathbf{M}$ by Cholesky decomposition such that (Ralston, 1965)

$$\Delta=\mathbf{M}\mathbf{M}^T \qquad [21.4.1]$$

where Δ is the variance-covariance matrix for $\mathbf{a}_t=(a_{t1},a_{t2},\ldots,a_{tk})^T$ in [21.2.6].

2. Obtain the vectors of initial values $\mathbf{Z}_{p_i,i},\mathbf{a}_{q_i,i}$, $i=1,2,\ldots,k$. (See next subsection for the method used to calculate the initial values.)

3. Following the two steps given next, generate $\mathbf{a}_{p+1},\mathbf{a}_{p+2},\ldots,\mathbf{a}_N$ which is a sequence of $N-p$ vectors each of which has dimension k and is NID($\mathbf{0}$,Δ). As in [21.2.6], the $p=\max(p_1,p_2,\ldots,p_k)$.

   (i) Simulate $\mathbf{e}_{t1},\mathbf{e}_{t2},\ldots,\mathbf{e}_{tk}$, which is a sequence of $N-p$ vectors each of which has dimension k and is distributed as NID($\mathbf{0}$,$\mathbf{1}$) where $\mathbf{0}$ is a $k\times1$ vector consisting of k zeroes, $\mathbf{1}$ is diagonal matrix of dimension $k\times k$ having entries of unity along the main diagonal.

   (ii) Calculate

$$a_{ti}=\sum_{j=1}^{i}m_{ij}e_{tj} \qquad [21.4.2]$$

   for $i=1,2,\ldots,k$ and $t=p+1,\ldots,N$.

4. Obtain $\mathbf{Z}_{p+1},\mathbf{Z}_{p+2},\ldots,\mathbf{Z}_N$, where each vector of observations at a given time has k entries, by using

$$\begin{aligned}Z_{ti}&=\phi_{ii1}Z_{t-1,i}+\phi_{ii2}Z_{t-2,i}+\cdots+\phi_{iip_i}Z_{t-p_i,i}+a_{ti}\\&\quad-\theta_{ii1}a_{t-1,i}-\theta_{ii2}a_{t-2,i}-\cdots-\theta_{iiq_i}a_{t-q_i,i}\end{aligned} \qquad [21.4.3]$$

for $i=1,2,\ldots,k$ and $t=p+1,p+2,\ldots,N$.

5. If another series of length N is required return to step 2.

The above algorithm is described for simulating stationary series having no Box-Cox transformations. If the original set of series were differenced and also were transformed using Box-Cox transformations, the techniques of Sections 9.5 and 9.6, respectively, could be employed in conjunction with the algorithm of this section to obtain synthetic sequences in the original untransformed domain.

Calculation of the Initial Values

The joint distribution of $\mathbf{Z}_{p_i,i}$ and $\mathbf{a}_{q_i,i}$, $i = 1,2, \ldots, k$, is used to generate the starting values for the simulation algorithm for a CARMA model. As demonstrated by Camacho (1984, p. 59), the joint distribution of

$$\upsilon = (\mathbf{Z}_{p_1,1}, \mathbf{Z}_{p_2,2}, \ldots, \mathbf{Z}_{p_k,k}, \mathbf{a}_{q_1,1}, \mathbf{a}_{q_2,2}, \ldots, \mathbf{a}_{q_k,k})^T$$

is multivariate normal having a mean of zero and variance covariance matrix given by

$$V = \begin{bmatrix} \gamma_{gh}(i-j) & \sigma_{gh}\psi_g(i-j) \\ Symm & \Delta \otimes \mathbf{I}_{q \times q} \end{bmatrix} \qquad [21.4.4]$$

where

$$\gamma_{gh}(r) = \langle Z_{t,g} Z_{t+r,h} \rangle, \quad g,h = 1,2, \ldots, k$$

$$\phi_i(B) \cdot \psi_i(B) = \theta_i(B), \quad i = 1,2, \ldots, k \qquad [21.4.5]$$

Ansley (1980) and Kohn and Ansley (1982) provide an algorithm to obtain the theoretical autocovariance function of the general multivariate ARMA model. This algorithm could be employed to calculate the terms $\gamma_{gh}(i-j)$ in [21.4.4]. However, due to the diagonal structure of the CARMA model, Camacho (1984, p. 61-62), has developed a computationally efficient algorithm for the calculation of the theoretical autocovariance function of the CARMA model.

The following algorithm can be used to obtain the initial values required in step 2 of the overall algorithm for simulating using the CARMA model given in Section 21.4.2.

1. Calculate $\psi_g(s)$, $g = 1, \ldots, k$; $s = 0,1, \ldots, \max\{p,q\}$ from [21.4.5].

2. Calculate the theoretical autocovariance functions $\gamma_{gh}(r)$,

   $$r = 1-p, \ldots, 0, \ldots, p-1, \quad 1 < g < h < k$$

3. Form the variance-covariance matrix $\mathbf{V}$ of υ given by [21.4.4] and obtain the lower triangular matrix $\mathbf{L}$ by Cholesky decomposition such that

   $$\mathbf{V} = \mathbf{L}\mathbf{L}^T$$

4. Generate $e_1 e_2, \ldots, e_{k(p+q)}$, a sequence of $k(p+q)$ NID (0,1) random variables and determine the vector of initial values by:

   $$U_j = \sum_{i=1}^{j} l_{ji} e_i, \quad j = 1,2, \ldots, k(p+q)$$

   Note that if another series is required only step 4 is needed.

21.5 PRACTICAL APPLICATIONS

21.5.1 Introduction

In order to clearly demonstrate the usefulness of CARMA modelling in water resources and environmental engineering, three case studies are presented. The first and third applications involve water quantity data while the second one deals with water quality time series. All three applications show how the model construction techniques of Section 21.3 can be conveniently used in practice to obtain models that adequately describe the series and possess efficient parameter estimates. In the third example where one series has more data points than the other, the estimation algorithm of Appendix A21.1 is employed so that all of the measurements can be used for efficiently estimating the CARMA model parameters. These three applications were originally presented by Camacho et al. (1985).

21.5.2 Fox and Wolf Rivers

Average annual riverflows in m^3/s for the Fox River near Berlin, Wisconsin, and the Wolf River near London, Wisconsin, are available from Yevjevich (1963) and also the hydrological data tapes of Colorado State University at Fort Collins, for the years from 1899 to 1965. A plot of the data is given in Figure 21.5.1, where the overall shapes and dependencies of the data can be compared. In order to facilitate these comparisons, the y-axes have been purposely deleted and the data have been scaled so that each of the two series takes up half the graph (these considerations were also taken into account to produce the plots given in Figures 21.5.3 and 21.5.5).

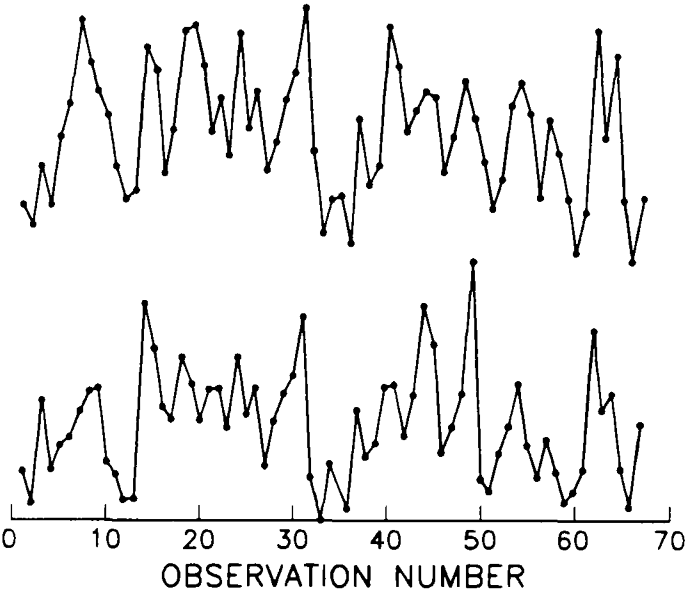


Figure 21.5.1. Annual riverflows for the Fox and Wolf Rivers (m^3/s).

Because the Fox and Wolf Rivers lie within the same geographical and climatic region of North America, a priori one may expect from a physical viewpoint that a CARMA model would be more appropriate to use than separate univariate ARMA models. Subsequent to taking a natural logarithmic transformation of the observations in both time series, univariate identification results from Chapter 5 suggest that it may be adequate to fit a MA model of order one (i.e., MA(1)) given in [3.3.1] to each data set. After prewhitening each series using the calibrated MA(1) model, the residual CCF for each series is calculated using [21.3.1] with the prewhitened

Fox and Wolf riverflows in order to obtain the graph of the residual CCF shown in Figure 21.5.2, along with the 95% confidence limits. Because the residual CCF in this figure is only significantly different from zero at lag zero, this indicates that a CARMA model could be fitted to the logarithms of the bivariate series. Additionally, the fact that each series can adequately be described by a univariate MA(1) model suggests that the following CARMA(0,1) model should be used.

$$\log Z_{ti} - \mu_i = (1 - \theta_{ii1})a_{ti}, \quad i = 1,2 \qquad [21.5.1]$$

where $i = 1$, and $i = 2$ refer to the Fox and Wolf logarithmic riverflows, respectively, μ_i is the theoretical mean of the logarithmic series for Z_{ti}, and the general definitions of all parameters and variables follow [21.2.7].

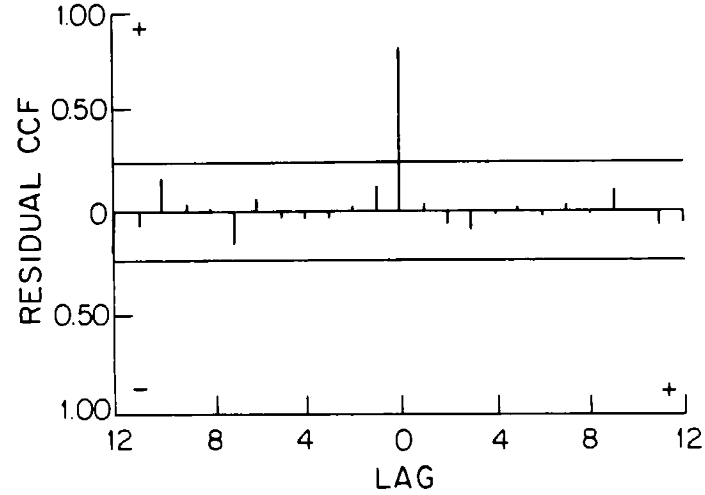


Figure 21.5.2. Residual CCF for the Fox and Wolf Rivers.

Table 21.5.1 lists th'e parameter estimates along with their standard errors appearing in brackets, using the univariate approach (Appendix A6.1) and the joint estimation algorithm described in Section 21.3.3. As can be observed in Table 21.5.1, there is a significant reduction in the variance of the parameter estimates when the joint estimation is employed. This in turn means that the relative efficiency of the univariate estimates with respect to the joint multivariate estimator is much less than unity. This relative efficiency is calculated using

$$eff = var(\hat{\theta}_{ii1})/var(\bar{\theta}_{ii1}) \qquad [21.5.2]$$

where $\hat{\theta}_{ii1}$ and $\bar{\theta}_{ii1}$ are the joint and univariate estimates, respectively, for the parameter θ_{ii1}. The correlation between $\hat{a}_{t1}$ and $\hat{a}_{t2}$ is calculated to be 0.78. When the residuals of the CARMA(0,1) are subjected to residual checking, no misspecifications of the fitted model are detected.

Table 21.5.1 Parameter estimates for the CARMA model and univariate models for the Fox and Wolf Rivers.

	Fox River	Wolf River
Univariate	-0.483	-0.411
Estimates of θ_{ii1}	(0.110)	(0.111)
Joint	-0.626	-0.543
Estimates of θ_{ii1}	(0.075)	(0.080)
Efficiency of Univariate Estimator	0.465	0.519
Mean of Log Z_{ti}	3.39	3.84
	(0.037)	(0.042)
Residual Variance	5.52×10^{-2}	7.5×10^{-2}

21.5.3 Water Quality Series

In the second example, two series corresponding to different measurements of the concentration of nitrogen in the Middle Fork Creek near Seebe, located in the Province of Alberta, Canada, are modelled. The series represent monthly measurements of total nitrogen and nitrogen Kjeldahl from 1972 to 1979 and are part of an overall data set that are studied using both exploratory and confirmatory data analysis tools in Sections 22.3 and 22.4, respectively. The seasonal adjustment algorithm of Section 22.2 was used to obtain the monthly means of the series from data available at irregular time intervals. A plot of the estimated monthly series is given in Figure 21.5.3.

Following Chapter 5 and Section 12.3.2, univariate identification techniques suggest that an adequate model for describing the natural logarithms of the total nitrogen series, Z_{t1} is a seasonal $AR(1)_6$ model of the form

$$(1-\phi_{116}B^6)(logZ_{t1}-\mu_1)=a_{t1} \qquad [21.5.3]$$

where $B^6\log Z_{t1}=\log Z_{t-6,1}$. An appropriate model to fit to the nitrogen kjeldahl series, Z_{t2} is an AR(1) model of the form

$$(1-\phi_{221}B)(\log Z_{t2}-\mu_2)=a_{t2} \qquad [21.5.4]$$

The univariate estimated parameters and their SE's given in brackets are are listed in Table 21.5.2.

A perusal of the residual CCF for the fitted models from [21.5.3] and [21.5.4], shows that only the CCF at lag zero is significantly different from zero. This identification result implies that a CARMA model is appropriate for fitting to the bivariate series. The specific parameters required in the two component equations of the overall CARMA model are the same as those

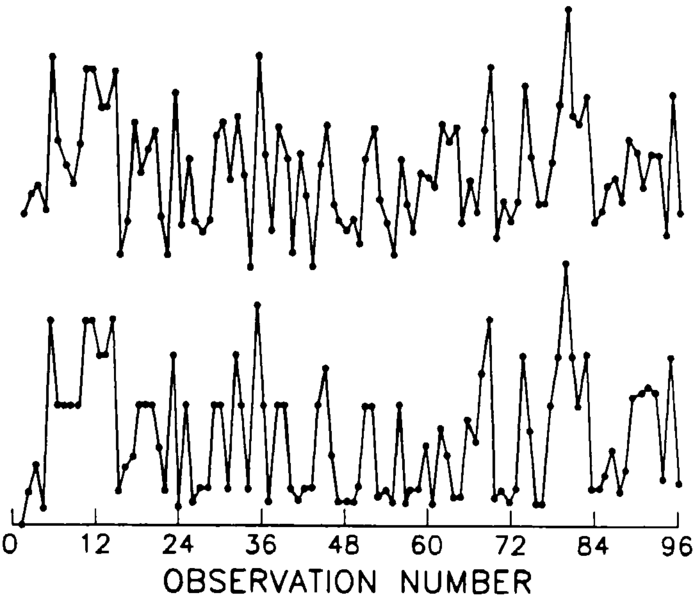


Figure 21.5.3. Concentration of total nitrogen and nitrogen kjeldahl (mg/*l*) for the Middle Fork Creek, near Seebe, Alberta, Canada.

used in [21.5.3] and [21.5.4]. Following the joint estimation procedure of Section 21.3.3, the efficient estimators for the CARMA model are calculated and displayed in Table 21.5.2. The reduction in the variances of the joint estimators compared with the variances of the univariate estimators is quite substantial. If only univariate series were used to estimate the parameters of the model for each one of the series, it would be necessary to increase the sample size of the series by a factor of four in order to obtain the same reduction in the variances of the parameters estimates. This increase in the sample size of the series is very expensive and in some cases infeasible. Consequently, this demonstrates that the CARMA model could also be employed to increase the accuracy of the parameters of the univariate models. The correlation at lag zero between $\hat{a}_{t1}$ and $\hat{a}_{t2}$ for the models given in [21.5.3] and [21.5.4], respectively, is found to be 0.88.

21.5.4 Two Riverflow Series Having Unequal Sample Sizes

As an example of two riverflow time series possessing unequal numbers of observations, consider the French Broad River at Asheville, North Carolina and the French Broad River near Newport, Tennessee, which have average annual flows from 1896 to 1965 and 1921 to 1965, respectively. As is the case for the application in Section 21.5.2, these flows are available from Yevjevich (1963) and also the hydrological data tapes of Colorado State University.

A plot of the 70 observations of the flows at Asheville and the 45 observations of the flows near Newport are displayed in Figure 21.5.5. Univariate MA(1) models like the ones in [21.5.1] were found to be adequate to fit the logarithms of the series. A plot of the residual CCF is given in Figure 21.5.6 Although the flows near Newport are measured downstream from the flows at

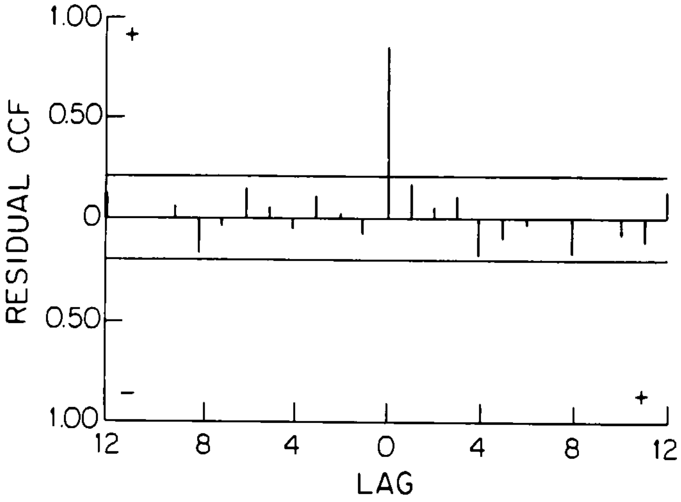


Figure 21.5.4. Residual CCF for the total nitrogen and nitrogen kjeldahl.

Table 21.5.2. Parameter estimates for the CARMA model and univariate models for the total nitrogen and nitrogen kjeldahl series for the Middle Fork Creek.

	Total Nitrogen	Nitrogen Kjeldahl
Univariate Estimates of	0.310	0.294
ϕ_{116} and ϕ_{221}	(0.097)	(0.097)
Joint Estimates of	0.141	0.141
ϕ_{116} and ϕ_{221}	(0.049)	(0.049)
Efficiency of Univariate Estimator	0.255	0.255
Mean of Log Z_{ti}	-1.33	-1.59
	(0.084)	(0.104)
Residual Variance	0.131	0.152

Asheville, implying that a TFN model (see Chapter 17) may be required to model the bivariate series, it is observed from the plot of the residual CCF that a CARMA model would suffice, (only the residual CCF at lag zero is significantly different from zero). This is due to the fact that annual riverflows are being considered and this ***temporal aggregation*** of the data, by its very nature, incorporates some of the lagged relationships, which would be expected to hold in the model of the system (Granger and Newbold, 1977). If monthly data or less temporal aggregated data were considered, a TFN model would probably be required to model the data. The algorithm given in Appendix A21.1 is used to jointly estimate the parameters of the model. These estimates are given in Table 21.5.3. The significant reductions in the variances of the estimators compared with the univariate estimates can be observed. The correlation at lag zero between the residuals of the two series is calculated to be 0.91.

21.6 CONCLUSIONS

As illustrated by the practical applications of the previous section, the CARMA family of models can be used to model efficiently hydrological and other types of environmental series. When taking the physical characteristics of the system being modelled into account along with output from the identification methods of Section 21.3.2, the CARMA class of models is often found to be the most appropriate type of multivariate model to use. The application of Section 21.5.4 shows that the CARMA model can be ideal for modelling time series formed by temporal aggregation. Another attractive feature of fitting this kind of model is that well developed, yet simple, model construction tools are currently available for use in practical applications. For example, when estimating the parameters of time series having equal and unequal sample sizes, the estimation procedures presented in Section 21.3.3 and Appendix A21.1, respectively, can be utilized. Furthermore, the flexible algorithm described in detail in Section 21.4.2 can be used for simulating synthetic sequences from a CARMA model.

Besides environmental series, the CARMA class of models has been successfully employed to model and forecast economic time series. Umashankar and Ledolter (1983), Moriarity and Salomon (1980) and Nelson (1976) used CARMA models to increase the efficiency of the estimated parameters and to improve the accuracy of the forecasts. Risager (1980) fitted CARMA models to mean annual ice core measurements. Research related to the development and application of CARMA models in hydrology was referred to throughout this chapter as well as Section 20.4.

Research in CARMA modelling can be extended in a variety of directions. For instance, as mentioned in Sections 21.3.5 and 20.3.3, model construction methods could be developed for various kinds of periodic CARMA models. Camacho (1984, Section 2.4) defines a contemporaneous TFN model in which the innovations among a set of k TFN models are contemporaneously correlated. If practical applications dictate the need for this rather sophisticated type of contemporaneous model, appropriate model construction methods could be developed.

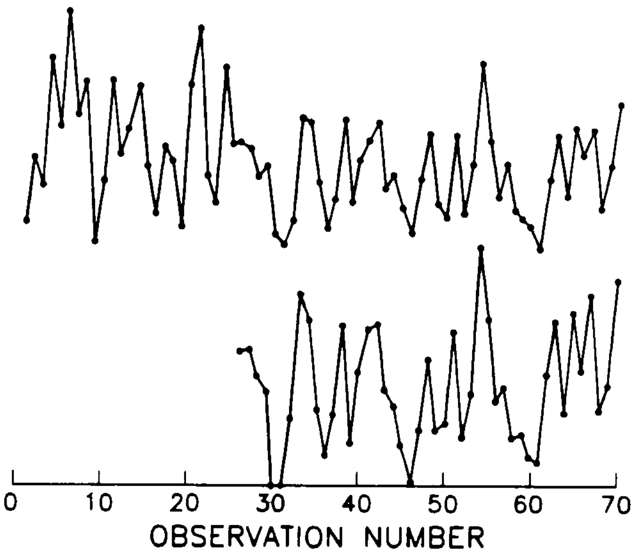


Figure 21.5.5. Annual riverflows for the French River at Asheville and near Newport (m^3/s).

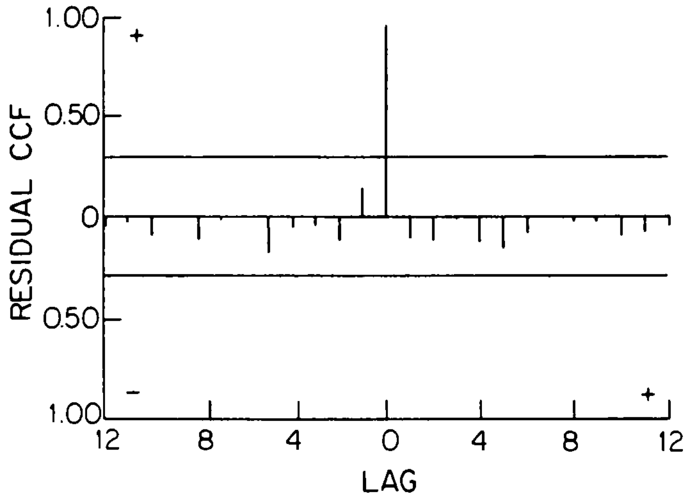


Figure 21.5.6. Residual CCF for the French Broad River at Asheville and near Newport.

Table 21.5.3. Parameter estimates for the CARMA model and univariate models for the French Broad River at Asheville and near Newport.

	At Asheville $n = 70$	Near Newport $n = 45$
Univariate Estimates of θ_{ii1}	-0.283 (0.115)	-0.469 (0.131)
Joint Estimates of θ_{ii1}	-0.170 (0.087)	-0.470 (0.081)
Efficiency of Univariate Estimator	0.572	0.382
Mean of Log Z_{ti}	4.01 (0.040)	4.36 (0.048)
Residual Variance	6.72×10^{-2}	5.79×10^{-2}

APPENDIX A21.1

ESTIMATOR FOR CARMA MODELS HAVING UNEQUAL SAMPLE SIZES

Within this appendix, an estimator is presented for obtaining maximum likelihood estimates for the parameters of a CARMA model [21.2.1] or [21.2.6] when the k time series used to calibrate the model do not have the same lengths. This algorithm was originally developed by Camacho et al. (1985). The CARMA estimator to be used with samples having the same number of observations over the same time period is given in Section 21.3.3.

When fitting models to multivariate hydrological data, it is common to find series with unequal numbers of observations. What is customary in this circumstance is to eliminate the additional information available in the longer series so that all the series end up with an equal number of observations. For example, Risager (1980) considered the modelling of a bivariate time series of mean annual ice core measurements for which data were available for the years 1861-1974 and 1169-1975, respectively. In his analysis, only data for the common period 1861-1974 could be used to jointly estimate the parameters of the model. Another possibility is to consider some of the observations of the shorter series as missing and use a procedure such as that given by Ansley and Kohn (1983) to estimate the parameters of the model. This approach, although sensible, is not computationally efficient for a large number of missing observations or for series having a large sample size. Another disadvantage of this procedure is the introduction of many additional parameters to be estimated, which reduce the accuracy of estimators. If a CARMA model is sufficient to fit to the data, the estimator described below can be employed for estimating the parameters of the CARMA model using all the available information in a very efficient way.

Suppose that the set of observations available for the series Z_{ti}, $i = 1,2,\ldots,k$, is given by

$$\{Z_{ti}\} = \{Z_{1-m_i,i},\ldots,Z_{0,i},Z_{1,i},\ldots,Z_{n,i}\} \quad \text{for } i = 1,2,\ldots,k,$$

where $t = 1-m_i, 1-m_i+1,\ldots,0,1,2,\ldots,n$, are the times at which the m_i+n observations in series i occur, $t = 1,2,\ldots,n$, are the common times for which all k series have measurements and hence n is the number of common observations across all k series. Although it is assumed that all the series go up to the same time n, it is possible to extend the procedures given below to include the case where not all the series end at the same time.

As in Section 21.3.3, let the parameters of the CARMA model for the ith series be contained in the vector

$$\beta_i = (\phi_{ii1},\phi_{ii2},\ldots,\phi_{iip_i},\theta_{ii1},\theta_{ii2},\ldots,\theta_{iiq_i})^T$$

Hence, the vector of parameters for the complete CARMA model is written as

$$\beta = (\beta_1,\beta_2,\ldots,\beta_k)^T$$

An approximate log-likelihood function of the CARMA model in [21.2.1] or [21.2.6] is

$$l(\beta,\delta) = -\frac{n}{2}\log\delta - \sum_{i=1}^{k}\frac{m_i}{2}\log\sigma_{ii} + S - \frac{1}{2}\sum_{i=1}^{k}\frac{S_{0i}}{\delta_{ii}}$$

where

$$S = \frac{1}{2}\sum_{t=1}^{n}\mathbf{a}_t^T\delta^{-1}\mathbf{a}_t$$

$$\mathbf{a}_t = (a_{t1},a_{t2},\ldots,a_{tk})^T$$

$$\delta = (\sigma_{ij}) \text{ and}$$

$$S_{i0} = \frac{1}{2}\sum_{t=1-m_i}^{0} a_{ti}^2$$

Using this approximation, it is possible to modify the algorithm given in Section 21.3.3 to estimate the parameters of a CARMA model when an equal number of observations are available for each series, to handle the case where the sample sizes are unequal.

The algorithm is as follows:

1. For each series Z_{ti}, $i = 1,2,\ldots,k$, obtain MLE's of the ARMA model parameters in [21.2.6] using an appropriate univariate ARMA estimation technique, such as one of those given by Newbold (1974), Ansley (1979), Ljung and Box (1979) or McLeod (1977) referred to in Section 6.2.3, with the complete set of observations $\{Z_{ti}\}$, $t = 1-m_i, 1-m_i+1,\ldots,0,1,\ldots,n$. Let the vector of univariate estimates be given by

$$\bar{\beta} = (\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k)^T$$

and set

$$\beta_0^0 = (\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_k)^T$$

2. Estimate $\delta = (\sigma_{ij})$ by solving the system of nonlinear equations

$$n\sigma_{ij} + \sum_{h=1}^{k} \frac{m_h}{\sigma_{hh}} \sigma_{ih}\sigma_{jh} = SS_{ij} + \sum_{h=1}^{k} \frac{S_{0h}}{\sigma_{hh}} \sigma_{ih}\sigma_{jh}$$

where

$$SS_{ij} = \sum_{t=1}^{n} a_{ti}a_{tj}$$

3. Calculate

$$\beta^{r+1} = \beta^r + \mathbf{V}(\delta l / \delta\beta)\Big|_{\beta=\beta^r}$$

where $\mathbf{V}$ is obtained as follows: Let

$$n\mathbf{I}_{gh} = plim(\partial^2 S / \partial\beta_q \partial\beta_h)$$

Then

$$\mathbf{V}^{-1} = [\sigma^{gh}\mathbf{I}_{gh}] + Diag\left[\frac{m_1}{n\sigma_{11}}\mathbf{I}_{11} + \frac{m_2}{n\sigma_{22}}\mathbf{I}_{22} + \cdots + \frac{m_k}{n\sigma_{kk}}\mathbf{I}_{kk}\right]$$

where $\delta^{-1} = (\sigma^{gh})$ and $Diag[\cdots]$ indicates a block diagonal matrix. The $[\sigma^{gh}\mathbf{I}_{gh}]$ can be determined using the algorithm provided by Ansley (1980) and Kohn and Ansley (1982). Iterations of the algorithm are continued until convergence is reached for giving the approximate MLE's of β. An application of the estimator in this appendix for fitting a CARMA model to two annual riverflow series having unequal sample sizes is furnished in Section 21.5.4.

PROBLEMS

21.1 In Sections 21.2.2 and 21.2.3, the subset and concatenation definitions are given for CARMA(p,q) models. For the following CARMA(p,q) models, write down the subset and concatenation definitions, the stationarity and invertibility conditions, and the entries in the variance covariance matrix for the innovations:

(a) CARMA(3,0)

(b) CARMA(0,4)

(c) CARMA(2,2)

(d) CARMA(4,3)

21.2 Select two annual time series that you think could be adequately modelled using a CARMA model. Follow the identification procedure of Section 21.3.2 to ascertain whether or not your supposition is justified.

21.3 Carry out the instructions of problem 21.2 for the situation when you have three time series.

21.4 Describe in detail how the estimation algorithm of Section 21.3.3 works for the following bivariate CARMA models:

(a) CARMA(1,0)

(b) CARMA(0,2)

(c) CARMA(1,1)

21.5 Find two annual time series that are only contemporaneously correlated with one another as indicated by the residual CCF. Fit a CARMA to these series and check that the calibrated model provides an adequate fit.

21.6 Carry out the instructions of problem 21.5 for the case when you have three time series.

21.7 Suppose that in a set of k seasonal time series, each time series has s seasons per year. Using both the subset and concatenation definitions of CARMA models from Sections 21.2.2 and 21.3.2, write down the equations for the periodic CARMA model.

21.8 Carry out the instructions of problem 21.7 for the case of a seasonal CARIMA model.

21.9 Suppose that you wish to simulate 10 values for a bivariate CARMA model. Using the algorithm of Section 21.4.2, explain in detail how these are calculated for the following bivariate CARMA models

(a) CARMA(1,0)

(b) CARMA(0,1)

(c) CARMA(1,1)

21.10 Select a CARMA which is of direct interest to you. After setting the model parameters at some reasonable values or else using a model that you have already calculated, simulate three synthetic series of lengths 100, 500 and 1,000. Now fit a CARMA model to each of these series. Compare your modelling results for the three sets of simulated sequences and draw appropriate conclusions.

21.11 Explain how you would calculate minimum mean square error forecasts for a CARMA model.

REFERENCES

CARMA MODELS

Camacho, F. (1984). Contemporaneous CARMA Modelling with Applications. Ph.D. thesis, Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1985). Contemporaneous autoregressive-moving average (CARMA) modelling in hydrology. *Water Resources Bulletin*, 21(4):709-720.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1986). Developments in multivariate ARMA modelling in hydrology. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, July 15-17, 1985, Fort Collins, Colorado. Engineering Research Center, Colorado State University, pages 178-197.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987a). Contemporaneous bivariate time series. *Biometrika*, 74(1):103-113.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987b). Multivariate contemporaneous ARMA models with hydrological applications. *Stochastic Hydrology and Hydraulics*, 1:141-154.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987c). The use and abuse of multivariate time series models in hydrology. In MacNeill, I. B. and Umphrey, G. J., editors, *Advances in the Statistical Sciences*, Festschrift in Honor of Prof. V. M. Joshi's 70th Birthday, Volume IV, Stochastic Hydrology, pages 27-44. D. Reidel, Dordrecht, The Netherlands.

Hipel, K. W. (1986). Stochastic research in multivariate analysis. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, Keynote Address, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium on Multivariate Analysis of Hydrologic Processes*, July 15-17, 1985, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

Jenkins, G. M. and Alavi, A. S. (1981). Some aspects of modeling and forecasting multivariate time series. *Journal of Time Series Analysis*, 2(1):1-47.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, Series B, 43(2):231-239.

McLeod, A. I. (1979). Distribution of the residual cross correlation in univariate ARMA time series models. *Journal of the American Statistical Association*, 74(368):849-855.

Risager, F. (1980). Simple correlated autoregressive process. *Scandinavian Journal of Statistics*, 7:49-60.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 32(2):121-133.

Salas, J. D., Tabios III, G. Q., and Bartolini, P. (1985). Approaches to multivariate modeling of water resources time series. *Water Resources Bulletin*, 21(4).

DATA SETS

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology paper no. 1, Colorado State University, Fort Collins, Colorado.

FORECASTING

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series.* Academic Press, New York.

Moriarty, M. and Salomon, G. (1980). Estimation and forecast performance of a multivariate time series model of sales. *Journal of Market Research,* 17:558-564.

Nelson, C. R. (1976). Gains in efficiency from joint estimation of systems of autoregressive-moving average processes. *Journal of Econometrics,* 4:331-348.

Umashankar, S. and Ledolter, J. (1983). Forecasting with diagonal multiple time series models: An extension of univariate models. *Journal of Market Research,* 20:58-63.

ESTIMATORS

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika,* 66(1):59-65.

Ansley, C. F. and Kohn, R. (1983). Exact likelihood of vector autoregressive-moving average process with missing or aggregated data. *Biometrika,* 70:275-278.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association,* 74(367):652-660.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. *Biometrika,* 66(2):265-270.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. *Biometrika,* 64(3):531-534.

McLeod, A. I. and Salas, P. R. H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 32:211-223.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika,* 61(3):423-426.

Nicholls, D. F. and Hall, A. D. (1979). The exact likelihood of multivariate autoregressive-moving average models. *Biometrika,* 66:259-264.

SIMULATION

Ansley, C. F. (1980). Computation of the theoretical autocovariance function for a vector ARMA process. *Journal of Statistical Computation and Simulation,* 12:15-24.

Kohn, F. and Ansley, C. F. (1982). A note on obtaining theoretical autocovariances of an ARMA process. *Journal of Statistical Computing and Simulation,* 15:273-283.

McLeod, A. I. and Hipel, K. W. (1978b). Simulation procedures for Box-Jenkins models. *Water Resources Research,* 14(5):969-975.

Ralston, A. (1965). *A First Course in Numerical Analysis*. McGraw-Hill, New York.

PART X

HANDLING MESSY ENVIRONMENTAL DATA

In an **environmental impact assessment** study, an analyst may be requested to detect and model trends in a data set provided by the client. Unfortunately, as discussed in detail in Section 19.3.1, there are many reasons as to why the quality of environmental data, such as a set of water quality time series, is often not very high. One of the major problems with environmental series is there are often missing data points among which there may be long periods of time for which no observations were taken. In addition, there may be one or more external interventions which affect the stochastic manner in which a series behaves. In other words, **environmental data are often quite messy.**

The major objective of Part X is to explain how an optimal amount of information from messy environmental series can be detected and modelled. To accomplish this, the data analysis methodology of Tukey (1977) can be followed. As initially mentioned in Section 1.2.4, the two main steps in an overall **data analysis** study are:

1. **Exploratory Data Analysis** (Section 22.3 as well as Sections 5.3.2, 19.2.3 and 24.2.2);
2. **Confirmatory Data Analysis** (Section 22.4 as well as Chapter 3 to 21 and Sections 23.3 and 24.2.3).

In Section 22.3, a range of useful exploratory data analysis tools are suggested for discovering important patterns and statistical characteristics such as trends, caused by external interventions. To demonstrate the insights that can be gained by employing exploratory data analysis tools, they are applied to water quality series in Sections 22.3 23.5, 24.2.2 and 24.3.2 within Part X, as well as many other locations in the book.

To rigorously characterize trends and other desired statistical traits which may be known in advance or else detected using exploratory data analysis studies (see Section 19.2.3), formal statistical techniques can be employed at the confirmatory data analysis stage. In Part X, the following three types of confirmatory data analysis methods are described and used in environmental applications:

1. **Intervention Analysis** (Section 22.4 and also Part VIII),
2. **Nonparametric Tests** (Chapter 23),
3. **Regression Analysis** (Section 24.2.3).

As explained in detail in Chapter 19 and exemplified by applications throughout Chapter 19 and in Section 22.4, **intervention analysis** can be used to ascertain the magnitudes of changes in the mean levels of a series due to one or more external interventions. To determine the most appropriate intervention model to fit to a given data set at the confirmatory data analysis stage, one can follow the identification, estimation and diagnostic check stages of model construction. However, in order to be able to fit an intervention model to the time series, a sequence of data points evenly spaced in time must be available. If there are missing observations, an appropriate **data filling** technique can be used to estimate an evenly spaced time series from the original observations which are available at irregular time intervals (see Section 19.3.2

for a discussion of data filling methods). For the situation where there is a large number of missing observations, a procedure based on **seasonal adjustment** can be used (see Section 22.2). As pointed out in Chapter 19, intervention analysis constitutes a very powerful technique for use in environmental impact studies.

An inherent advantage of using most nonparametric tests and regression analysis techniques given in Chapters 23 and 24, respectively, is that they can be used directly with evenly or unevenly spaced observations. As described in Chapter 23, since the early 1980's, a number of researchers have used **nonparametric tests** for detecting trends in water quality time series. A useful variety of nonparametric trend tests, including the Mann-Kendall and Spearman's partial rank correlation tests, are described in detail in Section 23.3. **Regression models** that can be employed as exploratory and confirmatory data analysis tools are discussed in Sections 24.2.2 and 24.2.3, respectively. The exploratory data analysis techniques described in Section 22.3 can, of course, be used in conjunction with the nonparametric tests and regression analysis, as well as the intervention models.

As summarized in Table 1.6.4, three **trend assessment methodologies** are presented in Part X for carrying out trend assessments of water quality and water quantity time series. For each of these studies, a methodological approach is developed within the overall framework of exploratory and confirmatory data analysis. The first study presented in Section 22.3 employs **intervention analysis** for modelling trends in water quality and water quantity time series measured in rivers. Within the second study discussed in Section 23.5, **nonparametric trend tests** and other appropriate statistical methods are utilized for discovering trends in water quality samples observed in a lake that may be affected by nearby industrial developments. Finally, in the third case study a methodology is designed in Section 24.3 for assessing trends in water quality time series measured in rivers. A particularly useful technique for tracing trends and accounting for the effects of flow upon a given water quality variable is the **robust locally weighted regression smooth** of Cleveland (1979) described in Section 24.2.2. Moreover, the **Spearman partial rank correlation test** is employed to detect trends in water quality time series when the impacts of seasonally are partialled out.

CHAPTER 22

EXPLORATORY DATA ANALYSIS AND INTERVENTION MODELLING IN CONFIRMATORY DATA ANALYSIS

22.1 INTRODUCTION

The main purpose of this chapter is to present a comprehensive methodology to identify and, if possible, stochastically model any trends which may be present in water quality as well as other kinds of environmental time series. These trends, if any, may be due to the presence of known or unknown interventions such as various types of land-use changes. In addition to possibly being affected by external interventions, usually a given water quality variable is measured at irregular time intervals and often there are large time gaps at which no data are collected. Consequently, water quality data are often very *messy* and systematic procedures are developed in this chapter, as well as Chapters 23 and 24, to optimize the amount of meaningful statistical information which can be gleaned from the currently available data.

As explained by Tukey (1977) and also briefly mentioned in Sections 1.2.4, 5.3.2, 19.2.3 and 24.2.2, there are usually two major steps in a statistical study. The first step is called *exploratory data analysis* and the objective of this phase of the work is to uncover important properties of the data by executing simple graphical and numerical studies. Some of the techniques available for this phase include a graph of the data against time, the 5-number summary graph which Tukey (1977, Ch. 2) calls the box-and-whisker plot, cross-correlation function, Tukey smoothing (Tukey, 1977, Ch. 7) and the autocorrelation function. The purpose of the next step which is referred to as *confirmatory data analysis* is to confirm statistically in a rigorous fashion the presence or absence of certain properties in the data. For example, when sufficient measurements have been taken for a water quality variable, exploratory data analysis may indicate that there is a possible trend in the data due to a known external intervention. Following this, the *intervention analysis* approach of Chapter 19 can be utilized as a confirmatory data analysis tool to determine if there has been a significant change in the mean level of the series.

The exploratory and confirmatory stages of data analysis can be compared to the process which takes place after a crime is committed (Tukey, 1977). At the exploratory stage of investigating a crime, a sleuth uses forensic tools and his common sense to discover evidence about the crime. If the detective does not understand how to execute an investigation, he may fail to look in the proper places for the criminal's fingerprints. On the other hand, if the investigator has no fingerprint powder he will not detect fingerprints on most objects. In an analogous fashion, the statistical analyst requires both the *tools of the trade* and *common sense*.

In the criminal justice system, the suspected criminal is taken to court after the collection of evidence by the investigative bodies. Following the evaluation of the available evidence, the jury and judge must ascertain if the criminal is guilty based upon the current information. Likewise, in a statistical study the purpose of the second main step, confirmatory data analysis, is to *verify quantitatively* if suspected statistical characteristics such as different kinds of trends are actually present in the data. When enough evidence is available, the results of a confirmatory data analysis can be quite useful to decision makers. For instance, when intervention analysis is employed in an environmental impact assessment study for properly confirming the presence of trends in water quality time series, the results can be used in court for forcing the polluters to adopt appropriate pollution abatement procedures.

Many exploratory and confirmatory methods require that equally spaced data be available, and as is pointed out earlier in this section, environmental series are often measured at uneven time intervals. Accordingly, in the next section a methodology based on *seasonal adjustment* is devised for estimating the entries of an average monthly time series when daily values are available at irregular time intervals and often there are time gaps spanning many months for which no measurements were taken. In addition to estimating values for a monthly sequence, this procedure can of course be used for estimating averages at other equal time intervals such as weekly or quarterly intervals by having fifty-two and four seasons per year, respectively.

Following the section on data filling, specific *exploratory data analysis techniques* are described in Section 22.3. In order to demonstrate clearly the efficacy of using exploratory data analysis and, when appropriate, the confirmatory data analysis tool of intervention analysis, *practical applications* are presented throughout the chapter. Possible *trends* in water quality and riverflow series are examined for two locations in Canada. In the province of Alberta, Canada, both exploratory and confirmatory data analysis techniques are employed to ascertain the effects of cutting down a forest upon total organic carbon and turbidity in the Cabin Creek near Seebe. On the Mill River near St. Anthony in Prince Edward Island, exploratory data analysis results suggest that perhaps due to acid rain, alkalinity levels may be increasing over time. These illustrative applications were originally presented in the paper by McLeod et al. (1983) and are in fact part of an extensive environmental study executed by the authors in which fifty environmental time series were exhaustively analyzed.

Besides the *intervention analysis* approach of Chapter 19 and Section 22.4 in this chapter, other confirmatory data analysis techniques include the *nonparametric tests* and *regression analysis* methods of Chapters 23 and 24, respectively. An advantage of most nonparametric tests and regression analyses is that they can be used with observations measured at either unequally or equally spaced time intervals. In fact, as pointed out in Chapter 23, nonparametric tests have been used extensively for checking for the presence of trends in water quality time series. In Section 23.3 and Appendices A23.1 to A23.3, many of these nonparametric tests are described in detail. As an alternative procedure, the regression models of Chapter 24, offer a promising parametric approach for modelling trends in unevenly spaced measurements.

It is important that the scientist always keep in mind the fact that *sufficient data* or information must be available if he or she wants to carry out confirmatory data analyses. Until further measurements are available, inadequate series must for the present time be "thrown out of court" due to lack of sufficient evidence. Often a detective has "a feeling" about whether or not a person is guilty of a crime. If he thinks that the suspect is actually guilty, he will continue to follow his prey until he collects sufficient information so the courts can eventually convict him. The same situation holds for statistical studies. Even though it may not be presently feasible to fit an intervention model or another type of confirmatory model to a specific water quality time series, if this series is deemed important, the further collection of data will eventually permit a full confirmatory data analysis study.

22.2 DATA FILLING USING SEASONAL ADJUSTMENT

Many exploratory data analysis methods are valid for use with either unequally or evenly spaced data. However, Tukey smoothing, which is explained in Section 22.3.5, is an example of an exploratory tool where the measurements, or estimates thereof, must be available at equal time intervals before the method should be used. Except for many of the nonparametric tests and also the regression analysis methods presented in Chapters 23 and 24, respectively, all of the stochastic models described in this book, including the intervention models of Chapter 19 and Section 22.4, can only be used with evenly spaced data at the confirmatory data analysis stage. Therefore, when data are unevenly spaced, procedures are required for creating an evenly spaced sequence which stochastically represents what could have occurred historically. As explained in Section 19.3, intervention analysis can be employed for estimating missing values from an evenly spaced data set when the number of unknown observations is not too large (usually not more than 5% of the data set). However, for evenly spaced daily observations with a large number of missing values, a different procedure must be adopted for estimating a sequence of evenly spaced average monthly values. The particular technique presented in this section is related to methods developed for seasonal adjustment models.

In *seasonal adjustment* modelling, a time series is decomposed into various components, one of which is the seasonal term. Various seasonal adjustment procedures are available and the reader may wish to refer to the statistical literature for a description of these techniques (see, for example, Kendall (1973), Shiskin et al. (1976), Granger (1980), Hillmer and Tiao (1985), and Cleveland et al. (1990)). Suppose that x_t represents an observation at time t either for the original time series or for some Box-Cox transformation of the given data. One reason for invoking the Box-Cox transformation in [3.4.30] is to cause data that are not normally distributed to approximately follow a normal distribution. For instance, a logarithmic transformation may reduce the skewness and improve the symmetry of the distribution if there are quite a few large values in the series. When the variance of a series depends on the level of the series, this transformation may rectify the problem. Furthermore, as explained in Section 3.4.5 and elsewhere, a Box-Cox transformation can often alleviate problems with the properties of the residuals of the stochastic model fitted to the series of equally spaced data.

An additive seasonal adjustment model can be written as:

$$x_t = C_t + S_t + I_t = C_r + S_m + I_t$$

where t is the Julian day number (i.e., the number of days since January 1, 4713 B.C.), r is the year, m is the month for monthly data, C_t or C_r is the trend factor for modelling relatively long term causes, S_t or S_m is a stable seasonal factor which is assumed not to evolve with time, I_t is the nonseasonal irregular component made up of short-run effects and is not necessarily white noise. The original seasonal adjustment algorithm presented by McLeod et al. (1983) consists of the following steps:

1. Obtain preliminary estimates of C_t, S_t and I_t. $\bar{C}_t = \bar{C}$ is taken to be a constant which is equal to the median of x_t. To get $\bar{S}_t$, first calculate $\bar{S}'_m$ as the median of $x_t - \bar{C}$ for the data in the mth month. Then use $\bar{S}_m = \bar{S}'_m - \frac{1}{12}\sum_{m=1}^{12}\bar{S}'_m$. Estimate the irregular component utilizing

$$\bar{I}_t = x_t - \bar{C} - \bar{S}_m$$

2. Replace far-out values in the $\bar{I}_t$ series by the nearest outer fence (see Section 22.3.3 on box-and-whisker graphs for definitions of far-out values and outer fences) to form the irregular series $\bar{I}'_t$. The process of replacing far-out values by outer fences is called *Winsorizing* (Tukey, 1977).

3. Estimate the deseasonalized series given by

$$D_t = \bar{C} + \bar{I}'_t$$

4. Determine the revised trend estimate, $\tilde{C}_t$, where each year in $\tilde{C}_t$ is the mean of D_t for that year. If no data are available for the rth year, the mean of D_t for surrounding years is used.

5. Calculate the revised seasonal component

$$\tilde{S}_m = \tilde{S}'_m - \frac{1}{12}\sum_{m=1}^{12}\tilde{S}'_m$$

where $\tilde{S}'_m$ is the median of $x_t - \bar{C}_t$.

6. The revised irregular series is estimated using

$$\tilde{I}_t = x_t - \tilde{S}_m - \tilde{C}_r$$

7. Winsorize the revised irregular series, $\tilde{I}_t$, to obtain the Winsorized series, $\tilde{I}'_t$. This is accomplished by replacing the far-out values of $\tilde{I}_t$ by the appropriate outer fences.

8. Obtain an adjusted version (i.e., Winsorized) of the x_t series using

$$x'_t = \tilde{C}_r + \tilde{S}_m + \tilde{I}'_t$$

For a given month for a specified year in which data were originally given, take the median of the x'_t values to get the estimated average monthly value.

9. Adjust the trend for each year by employing

$\tilde{C}_r = \tilde{C}_r$ + mean of $\tilde{I}'_t$ for the whole series.

10. To obtain an estimated monthly average value for a given month in which no data were given use

$$\bar{x}_{r,m} = \tilde{C}_r + \tilde{S}_m$$

where $\bar{x}_{r,m}$ is the estimated monthly value for the rth year and mth month. The total estimated monthly series is formed by using Steps 8 and 10. Note that if a Box-Cox transformation is taken of the given data, then an inverse Box-Cox transformation must be invoked to obtain the estimated monthly averages for the original untransformed series.

In Section 24.2.2, the *robust locally weighted regression smooth (RLWRS)* of Cleveland (1979) is explained as a flexible procedure for smoothing a time series. The above seasonal adjustment algorithm can be improved by employing the RLWRS in the algorithm. In particular, the fourth step in the algorithm becomes:

4. Determine the revised trend estimate, $\tilde{C}_t$, as the RLWRS fitted to the deseasonalized series, D_t.

In order to demonstrate how well the seasonal adjustment algorithm works, consider the flows in m^3/s of the Cabin Creek near Seebe in Alberta, Canada, from January, 1964, till December, 1979. A daily flow value has been measured for each day during this time period and for each month in a given year an average monthly value can be readily calculated. Because riverflow measurements are often highly skewed, it is advantageous to take natural logarithms of the data. In Figure 22.2.1, the natural logarithms of the actual average monthly values are marked with black circles for one particular four year interval. For exactly the same days on which observations are missing for the turbidity data in the Cabin Creek, the corresponding daily observations are removed from the flow data. Following this, the seasonal adjustment algorithm is employed to estimate the average monthly flows of the logarithmic daily data for the period from 1964 to 1979 and these estimated flows are marked by circles in Figure 22.2.1. It should be pointed out that for the turbidity series and hence the estimated flows, only about 8% of the data are used in the seasonal adjustment algorithm. In addition, there are many months during which no observations are available. However, as can be seen in Figure 22.2.1, the estimated values from the seasonal adjustment algorithm are reasonably close to the actual entries during this four year period and also the other years not shown in Figure 22.2.1.

As noted in Section 22.1, the seasonal adjustment algorithm can also be used for estimating averages at equal time intervals other than monthly spacings. For instance, it can be employed for determining average bimonthly and quarterly time series. Moreover, the RLWRS discussed in Section 24.2.2 can be employed for improving step 4 of algorithm. The reader may wish to refer to Section 24.3.2 for a description of how the RLWRS can assist in analyzing trends of messy water quality series measured in rivers.

22.3 EXPLORATORY DATA ANALYSIS

22.3.1 Introduction

A wide range of exploratory data analysis tools are available for detecting important statistical characteristics contained in a data set (see, for example, Tukey (1977), Velleman and Hoaglin (1981), Berthouex et al. (1981), Chambers et al. (1983), Cluis (1983), McLeod et al. (1983), Hoaglin et al. (1983), du Toit et al. (1986) and Ramsey (1988)). In Section 19.2.3, a number of exploratory procedures are suggested for detecting known and unknown interventions in a time series and some of these techniques are discussed in detail in this section. Indeed, all of the identification tools which are recommended for designing the different kinds of time series models discussed throughout the book can be considered as exploratory data analysis techniques for specifically deciding upon the parameters required in these models. For example, in Figure 19.2.4, the graphical methods used to detect known and unknown interventions and the identification techniques needed to design the intervention models, could both be considered as exploratory data analysis tools. Nevertheless, in this section exploratory tools are presented which do not assume that a specific type of stochastic model will be used at the confirmatory data analysis stage. Because, in some situations, confirmatory data analysis may not be warranted, due, for example, to a lack of sufficient data, confirmatory data analysis may not be executed subsequent to exploratory data analysis. However, when a confirmatory data analysis is carried out, many of the results from the exploratory data analysis stage may be used along with specially designed

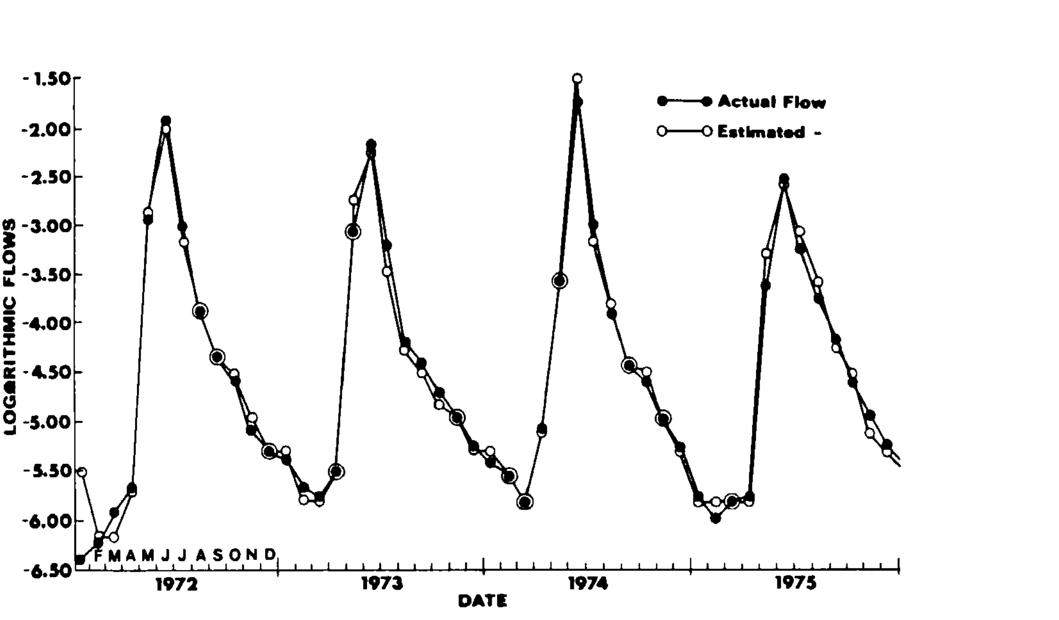


Figure 22.2.1. Monthly logarithmic flows of the Cabin Creek.

identification methods for constructing time series models at the confirmatory data analysis stage.

Some useful exploratory data analysis tools are presented in this section and the efficacy of using the techniques is demonstrated by applications to water quality and water quantity time series. Although many of the exploratory data analysis tools do not require the observations in a time series to be available at equally spaced time intervals, some of the techniques are designed for use with equally spaced measurements. The approaches discussed in this section which do not require equally spaced observations are:

1. graph of the data against time;
2. the 5-number summary graph which Tukey (1977, Ch. 2) calls a box-and-whisker plot;
3. cross-correlation function.

The following two techniques assume that the data points are separated by equal time intervals:

4. Tukey smoothing (Tukey, 1977, Ch. 7; Velleman and Hoaglin, 1981, Ch. 6);
5. autocorrelation function (ACF).

When data are unevenly spaced, an appropriate technique such as the seasonal adjustment algorithm from Section 22.2 or one of the other data filling methods of Section 19.2.3, can be employed for estimating the entries of an evenly spaced time series. Additionally, except for the third technique, all of the exploratory data analysis tools constitute valuable methods for detecting possible interventions.

The exploratory data analysis methods presented in this section and elsewhere can be employed for revealing interesting properties of the data under consideration. Each exploratory technique possesses its own inherent attributes that are useful for uncovering certain data characteristics. Because no single method can clearly portray everything there is to learn about the data, it is advantageous to examine the time series by employing a number of useful investigative graphical and numerical tools.

22.3.2 Time Series Plots

One of the simplest and more useful exploratory graphical tools is to plot the data against time. Characteristics of the data which may be easily discovered from a perusal of a graph include the detection of extreme values, trends, known and unknown interventions, dependencies among observations, seasonality, need for a data transformation, nonstationarity, and long term cycles.

When considering unequally spaced daily data, the actual time intervals between adjacent observations must be calculated before plotting the observations against time. A convenient technique to employ is to determine the *Julian day number* for each observation using the formula given by Hewlett-Packard (1977). With this information, the gap between adjacent observations can be determined as the difference of the Julian day numbers of the observations. This procedure is employed to obtain the graph in Figure 22.3.1 of the natural logarithms of the turbidity in the Cabin Creek, where each data point is marked by a small circle and is joined to its two neighbours by straight lines. As shown by the time gaps between observations, there are many days and even months during which no measurements were taken. For instance, from August 2 to November 22, 1975, inclusive, no observations were recorded.

Other examples of time series plots are presented throughout the textbook. In Chapter 2, Figure 2.3.1 displays the average annual flows of the St. Lawrence River at Ogdensburg, New York, for which there are no missing observations and no known interventions. An illustration of an annual series for which there is a known intervention is the average annual flows of the Nile River at Aswan in Figure 19.2.1. As seen in Figure 19.2.1 and discussed in Section 19.2.4, the completion of the Aswan dam in 1902 caused the average annual flows of the Nile River to decrease significantly from 1902 onwards. In Chapter 4, Figures 4.3.8, 4.3.10 and 4.3.15, show trends present in annual water use, electricity consumption, and Beveridge wheat price index time series, respectively. As noted in Section 4.3.3, each of these three nonstationary series can be adequately modelled using an ARIMA model.

Some interesting seasonal time series are also presented in the book. Consider, for example, Figure 1.1.1 or Figure 19.1.1 which displays the 72 average monthly phosphorous data (mg/l) from January 1972, until December, 1977, for measurements taken downstream from the Guelph sewage treatment plant located on the Speed River in the Grand River Basin, Ontario, Canada. As can be clearly seen, the commencement of phosphorous removal at upstream sewage treatment plants dramatically decreased the mean level of the series after the intervention date. Additionally, as indicated by the blackened circles, there are missing observations before and after the intervention. In Section 19.4.5, an intervention model is fitted to the water quality series in Figure 1.1.1 or 19.1.1 in order to statistically ascertain the effects of the intervention upon the level of the series and to estimate the missing observations.

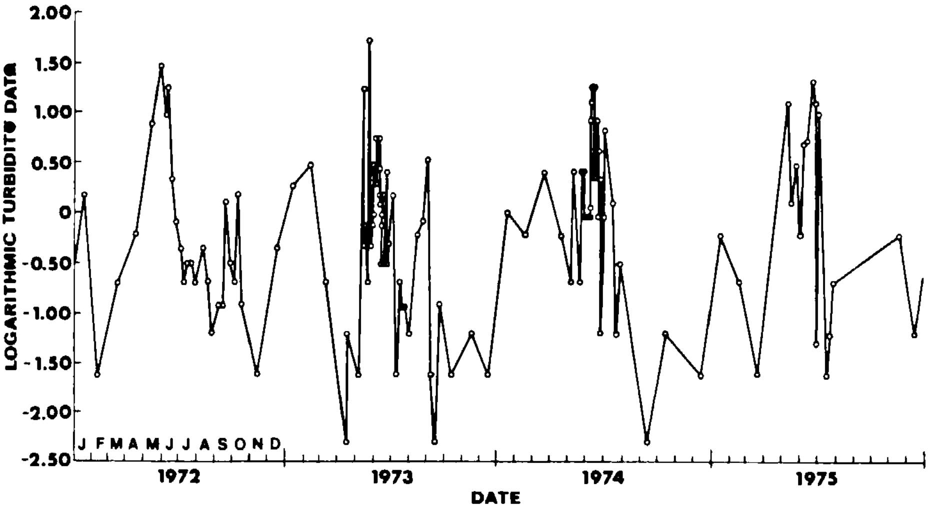


Figure 22.3.1. Natural logarithms of the turbidity (mg/l) data for the Cabin Creek.

22.3.3 Box-and-Whisker Graphs

A box-and-whisker graph is based upon what is called the *5-number summary* (Tukey, 1977, Ch. 2). For a given data set, the 5-number summary consists of the smallest and largest values, the median, and the 0.25 and 0.75 quantiles which are called *hinges*. When the data are ranked from the smallest to largest value, the first data point is the smallest value while the last entry is the largest value.

In order to calculate the values of *quantiles*, it is convenient to employ the operational definition of quantiles given by Chambers et al. (1983). Suppose that the given data represented by x_i for $i = 1,2,\ldots,n$, are ordered from smallest to largest such that the sorted data are denoted by $x_{(i)}$, $i = 1,2,\ldots,n$. If p represents any fraction between 0 and 1, the corresponding quantile is given by $Q(p)$. Whenever p is one of the fractions

$$p_i = (i - 0.5)/n \quad \text{for } i = 1,2,\ldots,n \qquad [22.3.1]$$

$Q(p)$ is assigned the value $x_{(i)}$, which is one of the given data points. For instance, if there were 10 observations, $x_{(2)}$ would have a p_i value of

$$p_2 = (2 - 0.5)/10 = 0.15$$

Hence, the 0.15 quantile, Q(0.15), would be exactly equal to $x_{(2)}$. When p is a fraction f of the way from p_i to p_{i+1}, one must use linear interpolation to estimate $Q(p)$. In particular, for this situation the interpolated quantile is calculated as

$$Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1}) \qquad [22.3.2]$$

Returning to the example for which there are 10 data points, the p_i for $x_{(3)}$ and $x_{(4)}$ are determined from [22.3.1] to be, respectively,

$$p_3 = (3 - 0.5)/10 = 0.25$$

$$p_4 = (4 - 0.5)/10 = 0.35$$

By utilizing [22.3.2], the quantile for $p = 0.31$ is determined to be

$$Q(0.31) = (1 - 0.6)Q(0.25) + 0.6Q(0.35)$$
$$= 0.4x_{(3)} + 0.6x_{(4)}$$

A plotting position is a value at which an ordered observation in a sample should be plotted for use on probability paper. In [22.3.1], p_i stands for the plotting position of the ith ordered value denoted by $x_{(i)}$. The plotting position given by [22.3.1] is actually a special case of the general plotting position formula written as

$$p_i = (i - \alpha)/(n + 1 - 2\alpha) \qquad [22.3.3]$$

where α is usually assigned values between 0 and 0.5. Detailed discussions regarding probability plotting positions are given by Barnett (1975) and Cunnane (1978) within the statistical and hydrological literature, respectively. As explained by these authors, when $\alpha = 0$ in [22.3.3] one obtains Weibull's formula which is recommended for use with uniformly distributed data. For normal observations, Blom's formula using $\alpha = 3/8$ in [22.3.3] should be employed. Finally, [22.3.1] is referred to as Hazen's formula and to obtain this, one substitutes $\alpha = 0.5$ into [22.3.3].

As noted by Chambers et al. (1983), there are many reasons for choosing p_i to be $(i - 0.5)/n$ in [22.3.1] or, equivalently, $\alpha = 0.5$ in [22.3.3], rather than some other value such as i/n. One consideration is that when the ordered observations are split into two groups exactly on an observation, the use of $(i - 0.5)/n$ means that the observation is counted as being half in the lower group and half in the upper group.

Because extrapolation must be done only when necessary and with great care, the formula in [22.3.2] for calculating $Q(p)$ should not be used outside the range of the data for which p is smaller than $0.5/n$ or larger than $1 - 0.5/n$. The safest rule for extrapolation is to define $Q(p) = x_{(1)}$ for $p < p_1$ and $Q(p) = x_{(n)}$ for $p > p_n$. According to this rule, $Q(0)$ and $Q(1)$ are assigned values of $x_{(1)}$ and $x_{(n)}$, respectively, which are the smallest and largest observations, respectively, in the given data set.

Equations [22.3.1] and [22.3.2] can be used with a data set of length $n \geq 2$. Keeping in mind that the only difference between a *percentile* and a quantile is that a percentile refers to a percentage of a data set and a quantile refers to a fraction of the data, these equations can also be used to calculate a percentile.

The *median*, given by $Q(0.5)$, divides the data into two groups of equal size. If n is odd, the median is $x_{((n+1)/2)}$. When n is even, $Q(0.5)$ is calculated using [22.3.2] as the average of $x_{(n/2)}$ and $x_{(n/2+1)}$, which are the two ordered values closest to the middle.

The lower and upper quartiles which are defined as $Q(0.25)$ and $Q(0.75)$, respectively, are called *hinges* by Tukey (1977). The distance between the first and third quantile, given by $Q(0.75) - Q(0.25)$ is called the *interquartile range*. This distance, which can be used to judge the spread of the data, is referred to by Tukey (1977) as the *H spread*.

To assist in characterizing extreme values, Tukey (1977) has suggested the following definitions. A *step* is 1.5 times the H-spread or interquartile range. *Inner fences* are one step outside hinges and *outer fences* are two steps outside hinges. Values between an inner fence and its neighbouring outer fence are called *outside*. Values beyond outer fences are *far-out*. Assuming that the data follow a given distribution, such as a normal distribution, one can calculate the expected numbers of outside and also far-out values, and compare these to the observed numbers.

When entertaining *seasonal data* such as monthly or quarterly data, it is instructive to calculate a 5-number summary plus outside and far-out values for each season. A convenient manner in which to display this information is to plot a box-and-whisker diagram for each season. Figure 22.3.2 depicts the monthly box-and-whisker plots for turbidity in the Cabin Creek before July 1, 1974, when part of the forest was cut down. In this figure, the data have not been transformed using a Box-Cox transformation. The upper and lower ends of a rectangle for a given month represent the two hinges and the thick line drawn horizontally within the rectangle is the value of the median. The minimum and maximum values for a particular month are the end points of the lines or *whiskers* attached to the rectangle or *box*. The far-out values are indicated by a circle in Figure 22.3.2, where far-out values are not marked if there are four or less data points for a given month. Below each month is a number which gives the number of data points used to calculate the box-and-whisker graph above the month. When there are not many data points used to determine a box-and-whisker plot for a given month, any peculiarities in the plot should be cautiously considered. The total number of observations across all the months is listed below November and December.

Another way to investigate *extreme values* is to calculate far-out values when all of the data across all of the seasons are used. Certainly, if a data point is far-out overall, the scientist should determine whether the measurement is accurate and represents what actually occurred or the observation is really due to measurement error or some other type of mistake. If the validity of a far-out overall or perhaps a far-out seasonal value is in doubt, in certain situations it may be advantageous not to include this data point in subsequent analyses. In the data filling procedure described in Section 22.2, far-out values are adjusted using a technique called *Winsorizing* (Tukey, 1977).

In addition to detecting far-out values, box-and-whisker diagrams have other uses. If the data are approximately symmetrical with respect to the median, they may follow a *symmetric distribution* such as the normal distribution. If there is an obvious lack of symmetry in a box-and-whisker graph, this may indicate the need for a transformation, such as the Box-Cox transformation in [3.4.30] to cause the data to be approximately normally distributed. Since, by definition, 25% of the data is contained between the median and a hinge, for normally distributed data the hinge is located 0.68 times the standard deviation from the median or mean. It can also be shown for normal data that inner fences are located a distance of 2.70 standard deviations on either side of the mean and the outer fences are a distance of 4.72 standard deviations from the mean. Consequently, the probability of having a far-out value with normally distributed data, is extremely small. Therefore, using a transformation to normalize a given data set will tend to

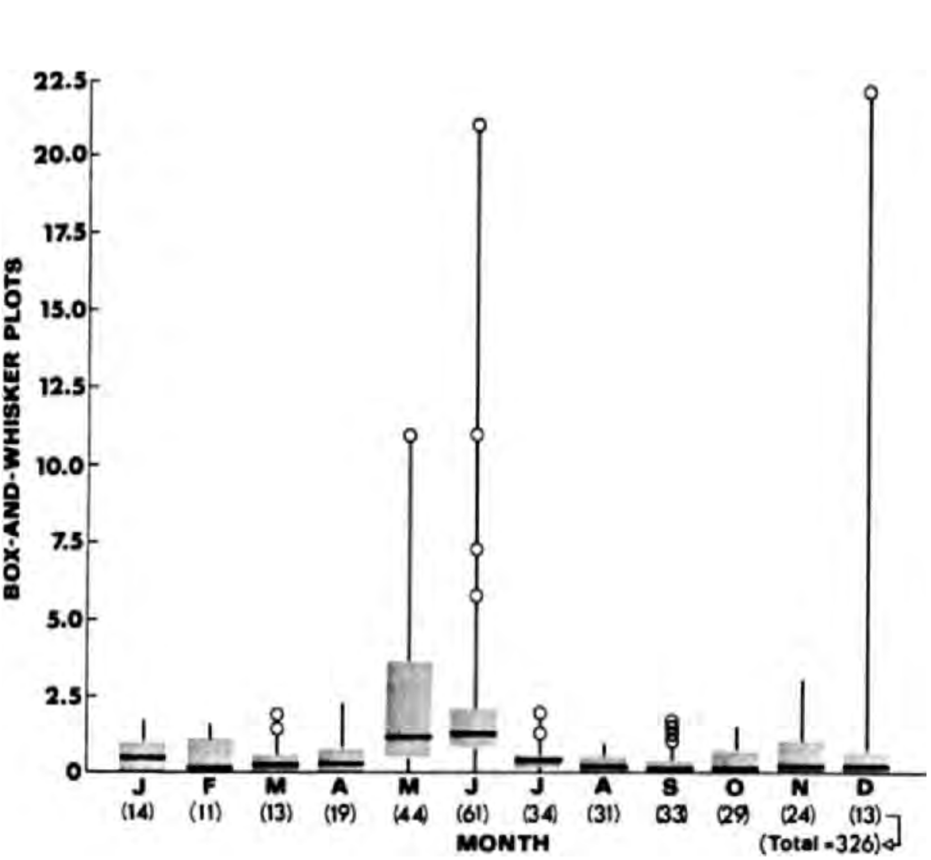


Figure 22.3.2. Box-and-whisker plots for turbidity (mg/l) in the Cabin Creek before July 1, 1974, where there is no data transformation.

reduce the number of far-out values.

For a given season in a box-and-whisker diagram, *symmetric data* would cause the median to lie in the middle of the rectangle and the lengths of the upper and lower whiskers would be about the same. Notice in Figure 22.3.2 for the turbidity data that the whiskers are almost entirely above the rectangle for all of the months and for six of the months there are a total of 14 far-out values. This lack of symmetry can at least be partially rectified by transforming the given data using the Box-Cox transformation in [3.4.30]. By comparing Figure 22.3.2 to Figure 22.3.3 where natural logarithms are taken of the turbidity data, the improvement in symmetry can be clearly seen. Furthermore, the Box-Cox transformation has reduced the number of far-out entries from 14 in Figure 22.3.2 to three in Figure 22.3.3.

Box-and-whisker plots can be employed as an important exploratory data analysis tool in *intervention studies*. If the date of the intervention is known, box-and-whisker diagrams can be constructed for each season for the data before and after the time of intervention. These two

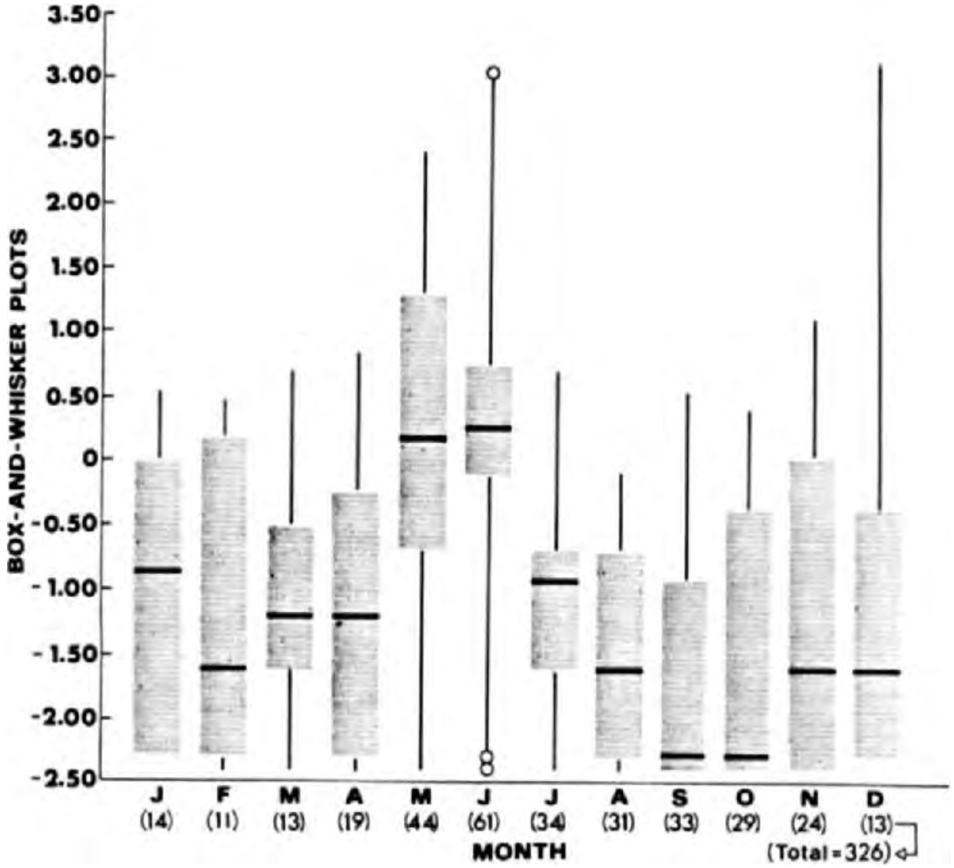


Figure 22.3.3. Box-and-whisker plots for turbidity (mg/l) in the Cabin Creek before July 1, 1974, where there is a logarithmic data transformation.

graphs can be compared to ascertain for which seasons the intervention has caused noticeable changes. When there are sufficient data, this type of information is crucial for designing a proper intervention model to fit the data at the confirmatory data analysis stage.

The Cabin Creek basin, which has an area of 2.12 km^2, was originally forested but from July to October, 1974, 40% of the forested area was clear-cut. Total organic carbon readings are available from March 17, 1971, to January 10, 1979. Figures 22.3.4 and 22.3.5 display the box-and-whisker plots of the natural logarithms of the total organic carbon in mg/l for the Cabin Creek before and after the intervention, respectively, caused by the removal of the trees. As can be observed, there are obvious drops in the medians for almost all the months after the intervention. These and other changes cannot be as easily detected in a plot of the entire series against time.

When examining seasonal box-and-whisker diagrams, such as those given in Figures 22.3.2 to 22.3.5, one may wish to compare the statistical characteristics of data among seasons in order to ascertain if there are any significant differences. One may be tempted, for instance, to check if two boxes do not overlap with one another which is the case for the June and July box-and-whisker plots in Figure 22.3.3. Unfortunately, the hinges which delineate the top and bottom of each box are not the appropriate guides to employ when checking for significant differences in a statistic such as the median between two seasons. A way for comparing medians between two box-and-whisker diagrams to check if they are significantly different, is to employ the *notched box-and-whisker plot* concept proposed by McGill et al. (1978). More particularly, for each box-and-whisker plot, a notch of specified size given below is drawn on the left and right side of the box with its centre at the median. When two box-and-whisker plots are compared to one another, the two medians are significantly different or not different according to whether the notches do not overlap or overlap, respectively. A suggested size for the notch is the

$$\text{median} \pm 1.58 \times \text{H--spread}/\sqrt{n}$$

where n is the number of data points used to construct the box-and-whisker plot for a given season. Assuming normality and independence of the data for each season, the significance level for which the median test is designed is approximately the 5% level. Examples of notched seasonal box-and-whisker plots are given in Figure 24.3.4 in Section 24.3.2.

Notice that at the bottom of each month for the seasonal box-and-whisker plots drawn in Figures 22.3.2 to 22.3.5, the number of data points is given. An approach for visually portraying the number of observations is to make the width of a box to be proportioned to the number of measurements. McGill et al. (1978) suggests drawing *variable-width box-and-whisker plot* for which the width is proportional to $\sqrt{n}$ for each season or group of data.

22.3.4 Cross-Correlation Function

In Section 16.2, it is explained how meaningful causality can be detected between two series labelled x_t and y_t, when the observations in each time series are equally spaced and sufficient observations are available. Subsequent to fitting an appropriate ARMA model to each of the series, the sample cross-correlation function (CCF) between the estimated residuals or prewhitened series of the two data sets can be calculated using [16.2.6]. By examining the properties of the residual CCF at negative, zero and positive lags, the type of causality between x_t and y_t can be ascertained. Illustrative applications for using this procedure are given in Section 16.3 and, in Sections 17.3.1, it is explained how the information from the residual CCF analysis can be used for identifying a TFN model to link the x_t and y_t series, when this type of model is warranted. In Sections 20.3.2 and 21.3.2, it is described how the residual CCF can be used to identify when a general multivariate ARMA model and a CARMA (contemporaneous ARMA) model are needed to model formally the mathematical relationship between x_t and y_t.

In exploratory data analysis, often there may be many missing data points and before two series can be prewhitened by fitting an ARMA model to each series, evenly spaced time series must be estimated. Further, at the exploratory data analysis stage one may only wish to have a general idea about the relationship between two series and to examine the CCF of the two given series at lag zero. If necessary, at a later stage in the data analysis study a proper residual CCF analysis can be executed. Consequently, even before an evenly spaced series is estimated for the

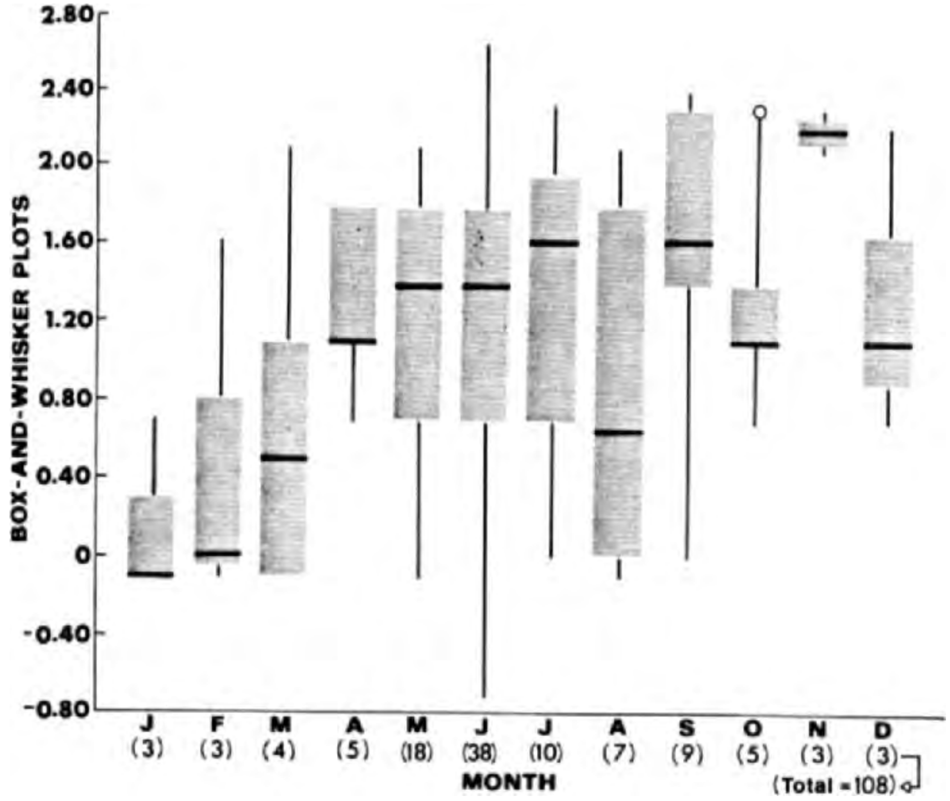


Figure 22.3.4. Box-and-whisker plots of the logarithmic total organic carbon (mg/l) in the Cabin Creek before July 1, 1974.

situation where data are missing, the CCF for x_t and y_t can be calculated for the values of the two series which are measured at the same time.

The CCF between two time series can be calculated to determine the amount of linear dependence between the two series. When x_t represents the observation recorded at time t for one series and y_t is the observed value at the same time for a second series, the *sample CCF* at lag zero can be calculated using

$$r_{xy}(0) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\left[\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2\right]^{1/2}} \qquad [22.3.4]$$

where n is the number of times observations occur at the same time in the two series, $\bar{x}$ is the mean of the x_t series, and $\bar{y}$ is the mean of the y_t series. The value of $r_{xy}(0)$ can range from -1 to

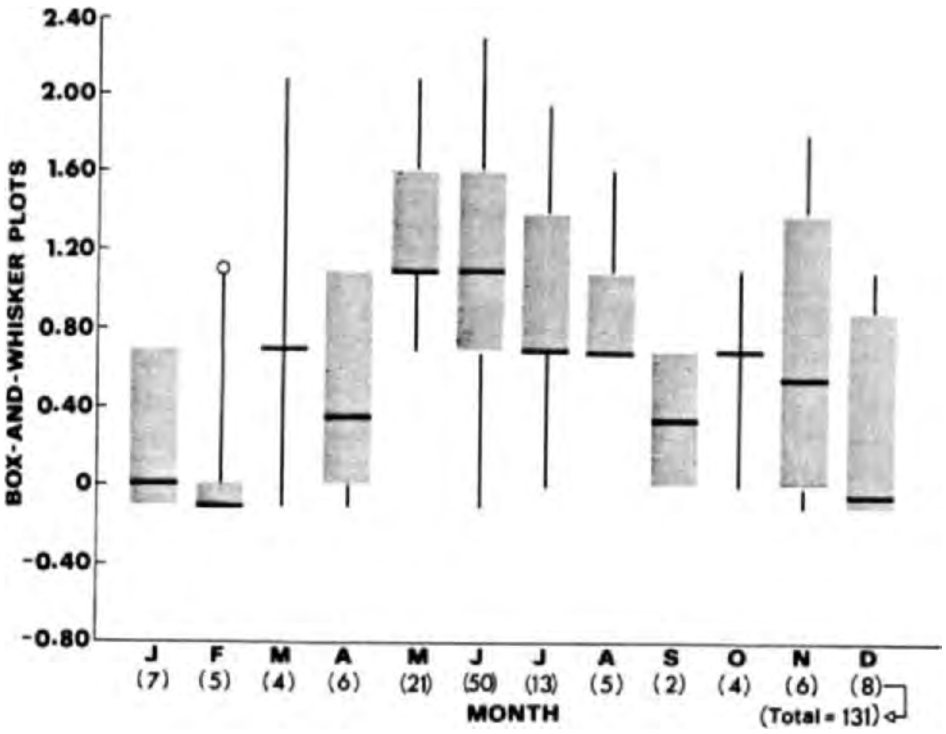


Figure 22.3.5. Box-and-whisker plots of the logarithmic total organic carbon (mg/l) in the Cabin Creek after October 31, 1974.

+1. If the x_t and y_t series are both white noise and also independent of one another, for large samples $r_{xy}(0)$ is normally independently distributed with a mean of zero and variance of $1/n$ (Haugh, 1976). Consequently, the 95% confidence limits are approximately $\pm 1.96n^{-1/2}$.

The sample CCF at lag zero can be calculated for either the original series or else the series transformed using the Box-Cox transformation in [3.4.30]. Recall that when the parameter $\lambda = 1$ in [3.4.30], this indicates that there is no transformation while $\lambda = 0$ means that each data point is transformed using natural logarithms. Suppose that a number of water quality variables plus riverflows have been measured at one location in a river. For a given site, $r_{xy}(0)$ in [22.3.4] can be calculated for all possible pairs of water quality and water quantity time series. Consider, for example, seven time series measured on the Mill River near St. Anthony, Prince Edward Island, Canada. Table 22.3.1 lists according to a number each of the seven series where the Box-Cox transformation in [3.4.30] used for each series is given. Below the list of series is the correlation matrix which is calculated using [22.3.4]. An (i,j) entry in the correlation matrix gives the value of the CCF at lag zero between series i and j which are defined above the correlation matrix in the table. For example, in Table 22.3.1 the CCF at lag zero between the 6th and 2nd series is

-0.817. Because this is the same as $r_{xy}(0)$ between series 2 and 6, the correlation matrix is symmetric and only the lower part of the matrix is presented. Moreover, the negative value indicates that the observations in one series tends to be larger whenever the values in the other series are smaller, and vice versa. Notice that all the diagonal entries have a value of unity since a series is fully correlated with itself at lag zero. If an $r_{xy}(0)$ value is within the range of $\pm 1.96n^{-1/2}$ it is automatically assigned a value of zero to indicate that it is not significantly different from zero.

Table 22.3.1. Cross-correlations for the Mill River time series.

DATA SETS

1. pH (pH Units), $\lambda = 1$
2. Stability Index, $\lambda = 0$
3. Daily Mean Discharge (m^3/s), $\lambda = 0$
4. Dissolved Sulphate (mg/l), $\lambda = 0$
5. Total Alkalinity (mg/l), $\lambda = 0$
6. Dissolved Calcium (mg/l), $\lambda = 1$
7. Water Temp. (degrees Celcius), $\lambda = 1$

CROSS-CORRELATION MATRIX

1.000						
-0.908	1.000					
-0.764	0.916	1.000				
0.268	-0.397	-0.436	1.000			
0.827	-0.965	-0.929	0.321	1.000		
0.535	-0.817	-0.793	0.000	0.768	1.000	
0.326	-0.376	-0.402	0.000	0.366	0.316	1.000

Measuring a large number of phenomena at a given location is usually quite expensive. If it is required to record less variables in order, for example, to allow the remaining items to be measured more frequently, the cross-correlation matrix may be helpful for deciding upon which variables to continue measuring. When one variable is highly correlated with another, then measuring only one of the variables furnishes an indication of the possible magnitudes of the actual values for the other unobserved variable. Consequently, based upon a firm understanding of the actual physical process plus the statistical evidence in the cross correlation matrix, it may be feasible to only continue to measure one of the series. If enough equally spaced measurements are taken of the remaining variable to permit the resulting time series to be thoroughly studied using a technique such as intervention analysis (see Chapter 19) at the confirmatory data analysis stage, this could be of great benefit to the decision makers.

A perusal of Table 22.3.1 reveals that many variables are highly correlated with one another. For example, notice in Table 22.3.1 for the Mill River time series that the stability index is highly correlated with pH($r_{2,1}(0) = -0.908$), daily mean discharge ($r_{3,2}(0) = 0.916$), total alkalinity ($r_{5,2}(0) = -0.965$), and dissolved calcium ($r_{6,2}(0) = -0.817$). Of course, it is known from a definition of the stability index that it is a function of the other mentioned water quality variables and this is confirmed by the appropriate entries in the correlation matrix in Table 22.3.1.

The zero entries in Table 22.3.1 demonstrate that sometimes there is no significant linear dependence between many of the variables. For example, in Table 22.3.1 for the Mill River series, the value for $r_{7,4}(0)$ between water temperature and dissolved sulphate is not significantly different from zero. When no significant cross-correlation exists between two series, then the decision about possibly dropping one of the series must be based upon other factors.

22.3.5 Tukey Smoothing

Introduction

Sometimes a graph of a given time series *blurs* statistical information in the data which a smoothed plot of the series at equally spaced time intervals may reveal more clearly. Consider, for example, Figure 22.3.6, which is a plot of the average annual total organic carbon in mg/l, for the Cabin Creek where the average annual entries are calculated using the estimated monthly values obtained from the seasonal adjustment algorithm developed in Section 22.2. In this graph, there appears to be a drop in the mean level of the series in the later years compared with the values in the early 1970's. When the *blurred smooth* in Figure 22.3.7 is studied, the general characteristics of the data are more clearly portrayed. Figure 22.3.7 is a blurred smoothed plot of the average annual total organic carbon for the Cabin Creek where the vertical lines reflect the magnitude of the rough or blur of the series and a smoothed observation is located at the midpoint of the bar. Notice from Figure 22.3.7 that the smoothing characteristics for the data before 1974 are more or less the same but from 1974 onwards there is an obvious decrease in the mean of the series. This property was also suggested by the box-and-whisker plots of the series shown before and after the intervention in Figures 22.3.4 and 22.3.5, respectively.

Although a *smoothed graph* does not contain any more information than what is already present in the plot of the raw data, in many instances the smoothed graph portrays the essential features much more clearly. The purpose of a smoothed curve is to reveal the systematic structure and interesting statistical characteristics of the data. Consider, for example, the blurred smoothed graph in Figure 22.3.8 for the total alkalinity in mg/l for the Mill River at St. Anthony in Prince Edward Island, Canada. This graph is a blurred smoothed plot of the average annual values which were calculated from the estimated monthly entries obtained from the seasonal adjustment algorithm in Section 22.2. In Figure 22.3.8, there is an obvious shift downwards in alkalinity from 1973 to 1977 followed by abrupt decreases in 1978 and 1979. Because the soil in the Mill River basin is sandy, acid rain could quickly drain through the ground without undergoing substantial chemical changes and thereby adversely affect the water quality. Consequently, the decrease in alkalinity in Figure 22.3.8 could be mainly due to acid rain which could severely affect the biological life in the river. However, it is still necessary to collect more data and determine when the acid rain intervention came into effect before proper confirmatory data analyses can be executed.

To construct a smoothed curve, consider qualitatively subdividing a given time series as

$$Data = Smooth + Rough$$

By filtering out the *rough* or noise portion of the data, the smoothed curve or *smooth* can be examined for important statistical features. The filter which maps the given series into a

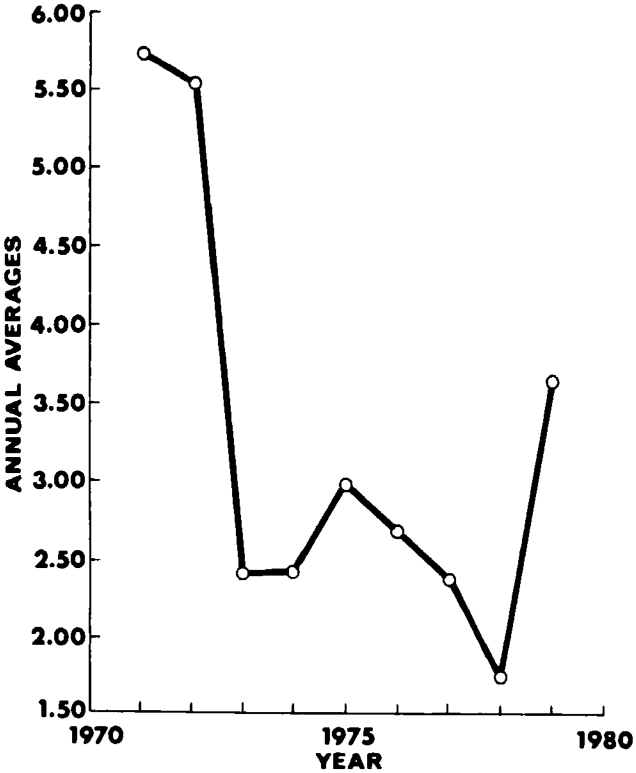


Figure 22.3.6. Estimated annual values of the total organic carbon (mg/l) in the Cabin Creek.

smoothed curve is referred to as a *smoother*. A trace plot of the smooth (against time) can display trends and changes in level of the series more clearly than a plot of the raw data. Of equal importance, a graph of the rough (over time) can reveal outliers, changes in variance and other unusual features.

Smoothed curves could be calculated for time series available at different time intervals between each pair of data points such as daily, monthly or yearly time separations. If one is attempting to detect *short term trends*, then it may be advantageous to examine smoothed curves and also time series plots for data points separated by short time intervals. However, at short time intervals long term trends may not be as easy to detect due to the large amount of rough in the data. Consequently, in order to discover *long term trends*, it may be advantageous to use annual data as is done in Figure 22.3.6 for the time series plot of the estimated annual values of total organic carbon (mg/l) in the Cabin Creek and Figure 22.3.7 of the blurred 3RSR smooth of the annual data plotted in Figure 22.3.6. As noted earlier, the long term trend in the total organic carbon time series can be easily visualized by examining these two figures. Within this section

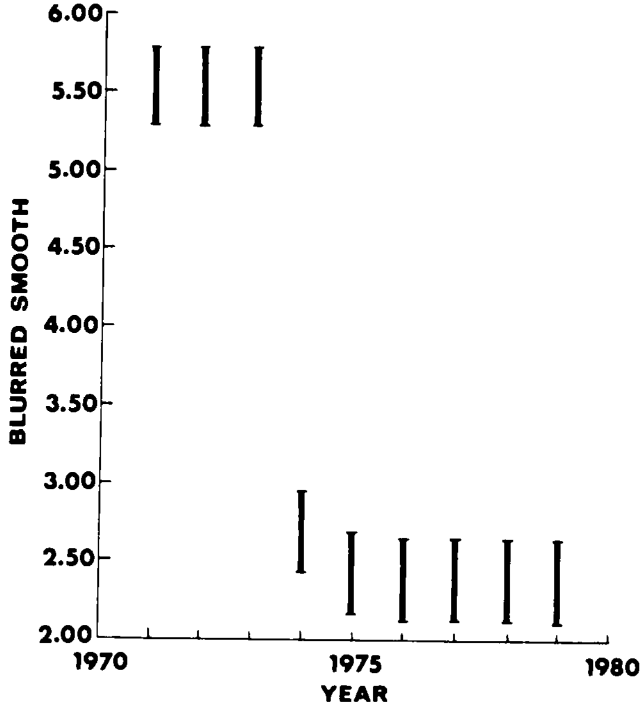


Figure 22.3.7. Blurred 3RSR smooth of the estimated average annual total organic carbon (mg/l) in the Cabin Creek.

and also Section 22.3.6, annual time series are considered so that long term trends can be conveniently discovered and their behaviour can be better understood.

The nonlinear smoothers developed by Tukey (1977, Ch. 7) and also discussed by McNeil (1977), are very flexible when used in practical applications and are capable of detecting all of the items discussed for a plot of the series except, possibly, for occasional outliers. Mallows (1980) explains the desirable properties that any smoother should possess and also presents some theoretical mathematical results for Tukey smoothers. Some of the more important attributes that a smoother should have include the ability to be responsive to abrupt changes in level, marginal distribution, and covariance structure.

Figures 22.3.7 and 22.3.8 are examples of what Tukey (1977, Ch. 7) calls a *blurred 3RSR smooth*. In fact, Tukey defines a variety of useful nonlinear smoothers. Within the next subsection the blurred 3RSR smooth is defined while the *4253H, twice, smooth* is described in the last part of Section 22.3.5. The reader may also wish to read about the flexible smooth of Cleveland (1979) which is based upon regression analysis and defined and applied in Sections 24.2.2 and 24.3.2, respectively. Finally, Velleman (1980) provides comparisons of robust nonlinear data smoothing algorithms.

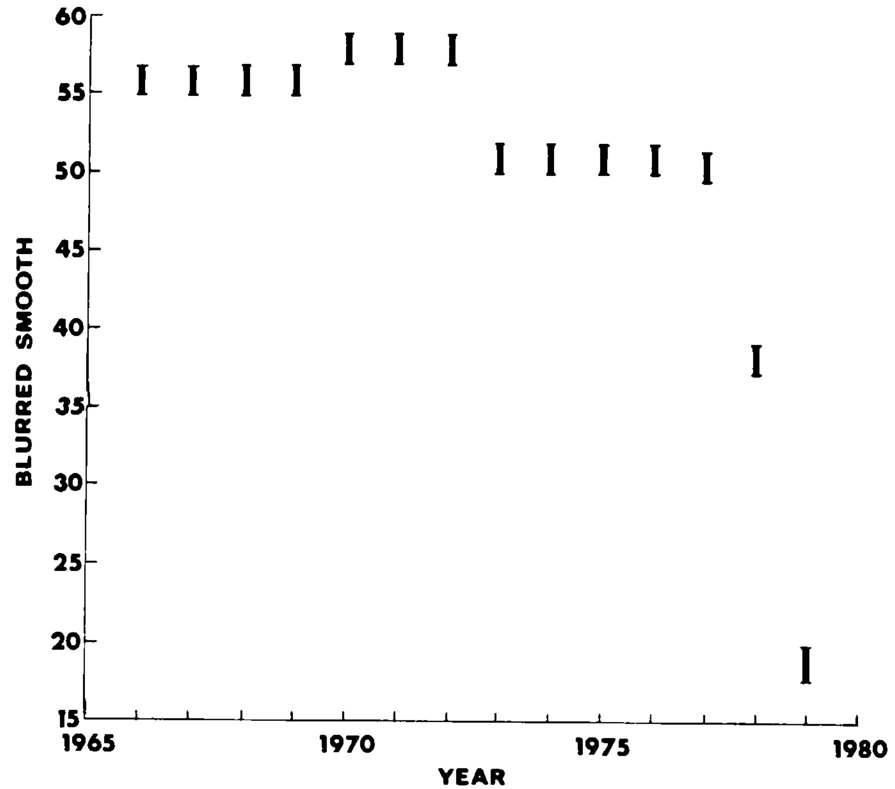


Figure 22.3.8. Blurred 3RSR smooth of the estimated average annual total alkalinity (mg/l) in the Mill River.

Blurred 3RSR Smooth

Blurred 3RSR smooths are displayed in Figures 22.3.7 and 22.3.8. When developing a *blurred 3RSR smooth* for a series $x_1, x_2, \ldots, x_n$, various calculations must be done (Tukey, 1977, Ch. 7; McNeil, 1977, Ch. 6) and these are outlined below.

1. *Smoothing by repeated medians of 3 (called 3R)* - To smooth the given time series using running medians of 3, replace the observation at time t by the median of x_{t-1}, x_t, and x_{t+1}. The smoothed values for the end points x_1 and x_n are calculated according to the rules in Step 2. Next, the smooth which was just determined is itself smoothed by using running medians of 3 and once again the smoothed end points are determined from Step 2. This procedure is repeated until the curve can be smoothed no further. The label 3R indicates that the smoothing is repeated using running medians of 3 until convergence is reached.

2. *Smoothing end points* - To obtain the smoothed values for the end points of the given series (or for the end points to any given sequence) replace x_1 by the median of x_1, x_2 and $3x_2 - 2x_3$ and substitute the median of x_n, x_{n-1}, and $3x_{n-1} - 2x_{n-2}$ for x_n.

3. *Repeated splitting (called SR)* - The use of medians instead of means tends to create mesas which are pairs of adjacent points with the same value that is below or above the points at each side of the mesa. Therefore, a *mesa* is simply a two-point local minimum or maximum. To smooth the mesas, split the series at the centre of each mesa and apply the end-point rule in Step 2 separately to the values on each side of each of the divisions. The resulting series is then smoothed using the 3R method in Step 1. The procedure of using the end-point rule for replacing the values at the mesas and then employing 3R is called splitting (i.e., S). If mesas still exist after splitting, the splitting is repeated until either all of the mesas disappear or convergence is reached. This is referred to as repeated splitting or simply SR. A computer program for calculating the 3RSR curve, which is created upon the completion of Step 3, is given by McNeil (1977, Ch. 6).

4. *Blurring* - When Steps 1 to 3 are used to obtain the smooth 3RSR a series of single points can be plotted. To reflect the variation, blur or rough in the series beyond the 3RSR curve, vertical bars can be plotted which are centered at each point on the 3RSR smooth. To calculate the length of the bars, first determine

   rough = data − smooth (i.e. 3RSR)

   for each point in the series. Wherever the rough has a value of exactly zero replace it by 0.5 as was suggested by Tukey (1977, Ch. 7). The bar length is then taken as the magnitude of the median of the absolute values of all the roughs. The 3RSR curve which is plotted using bars is called the blurred 3RSR smooth.

Summary - To obtain a blurred 3RSR smooth first determine the smooth 3R from Step 1 where the smoothed end points are calculated using Step 2. Next, determine the curve 3RSR by employing repeated splitting according to Step 3 and the relevant portions of Steps 1 and 2. Finally, Step 4 can be utilized to procure a blurred 3RSR smooth. Figures 22.3.7 and 22.3.8 are examples of blurred 3RSR smoothes which are calculated using the foregoing algorithm.

4253H, Twice Smooth

A particularly robust smooth is the *4253H, twice smooth* (Velleman and Hoaglin, 1981, Ch. 6). As indicated by the name, it involves taking medians of 4, then 2, then 5, then 3, then Hanning and then applying 4253H to the residuals of the first pass and adding this to the first pass smoother. To clarify how each step is calculated, an illustrative example is included in the explanation given below.

1. *Smooth using running medians of 4* - When determining the median of an even number of observations, the measurements being considered are ranked from smallest to largest and the median is taken as the average of the middle two values. Consequently, when calculating the median of four observations, the four values are listed in ascending order of magnitude, and the median is the average of the two middle numbers. For smoothing in using running medians of 4, which is simply called 4 smoothing, the endpoints themselves are just *copied on* in the 4 smooth series. The value located second from either end in the 4 smooth is simply the averages of the appropriate two end points in the given series. All

other entries in the 4 smooth are calculated as running median of 4.

As an illustrative example which is used to explain the entire 4253H, twice smooth, consider a hypothetical sequence of eight values given as:

5, 2, 4, 4, 0, 2, 3, 4.

The 4 smooth consisting of nine values is found to be:

5, 3.5, 4, 3, 3, 2.5, 2.5, 3.5, 4.

To explain how each value in the 4 smooth is determined, the calculations are given below for each of the numbers by starting on the left and working to the right.

5 = *copied on*

3.5 = *med*(5,2) = (5+2)/2 = 3.5

4 = *med*(5,2,4,4) = (4 + 4)/2 = 4

3 = *med*(2,4,4,0) = (2 + 4)/2 = 3

3 = *med*(4,4,0,2) = (2 + 4)/2 = 3

2.5 = *med*(4,0,2,3) = (2 + 3)/2 = 2.5

2.5 = *med*(0,2,3,4) = (2 + 3)/2 = 2.5

3.5 = *med*(3,4) = (3 + 4)/2 = 3.5

4 = *copied on*

2. *Smooth utilizing running medians of 2* - In step 1, except for the end points, each sequence of two numbers in the smooth can be thought of as lying on either side of the appropriate number in the original series in terms of the time axis. For example, the second and third entries from the left in a 4 smooth can be interpreted as residing on both sides of the second number in the original series. To line up in time the 4 smooth with the given series, a running median of two is applied to the 4 smooth after copying on the end points. In other words, the average is calculated for each sequential set of two values in the 4 smooth, excluding the two end points.

   In terms of the application, applying a 2 smooth to the previous 4 smooth creates the series of eight values given as:

   5, 3.75, 3.5, 3, 2.75, 2.5, 3, 4.

   Notice that the end points given by 5 and 4 are simply copied on. The second number from the left is calculated as:

   3.75 = *med*(3.5,4) = (3.5 + 4)/2.

3. *Smooth using running medians of 5* - When determining the median of an odd number sequence of observations, the measurements being entertained are ranked from smallest to largest and the median is selected as the value falling in the middle. For a 5 smooth, the ends are just copied on and lower order smoothing is employed near the ends while running medians of 5 are employed to calculate all other entries.

When the 5 smooth is applied to the smooth obtained at step 2, the result is

$$5, 3.75, 3.5, 3, 3, 3, 3, 4.$$

Once again the two end points are copied on as 5 and 4. The second entry from the left is determined as:

$$3.75 = med(5, 3.75, 3.5).$$

Likewise, the second value from the right is found to be:

$$3 = med(2.5, 3, 4).$$

All other values are determined using running medians of 5. For instance, the third entry from the left is found using:

$$3.5 = med(5, 3.75, 3.5, 3, 2.75).$$

4. *Smoothing employing running medians of 3* - Suppose that the series to be smoothed at this stage is represented as $x'_1, x'_2, \ldots, x'_n$, where the original series is represented as $x_1, x_2, \ldots, x_n$. The left and right end points of the 3 smooth are calculated as:

$$med(3x'_2 - 2x'_3, x'_1, x'_2) \text{ and}$$

$$med(x'_{n-1}, x'_n, 3x'_{n-1} - 2x'_{n-2}), \text{ respectively.}$$

Notice that the point $3x'_2 - 2x'_3$ corresponds to an estimate of x'_0 derived by extrapolating backwards the line joining $(2, x'_2)$ and $(3, x'_3)$. By working from left to right, all other entries, excluding the end points, are calculated as running medians of 3.

Applying the 3 smooth to the example as developed in the previous steps, produces the series

$$4.25, 3.75, 3.5, 3, 3, 3, 3, 3.$$

The left end point is determined using:

$$\begin{aligned} 4.25 &= med(3(3.75) - 2(3.5),5,4) \\ &= med(4.25,5,4). \end{aligned}$$

The second entry from the left is found as:

$$3.75 = med(5,3.75,3.5)$$

5. *Hanning* - Hanning (H) refers to a rather gentle smoothing operation for which a given value is replaced by a *running weighted average*. Let the current series to which hanning is to be applied be given as $(x_1'', x_2'', \ldots, x_n'')$. A practical running weighted average to employ for calculating the hanned value at time t is

$$\frac{1}{4}x_{t-1}'' + \frac{1}{2}x_t'' + \frac{1}{4}x_{t+1}''.$$

As can be seen, the weights given by $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ sum to unity. Although an unlimited number of running weight averages are available for use, the hanning smoother utilized here employs the weights given above. When employing hanning, the two end points are copied on and all other entries are calculated using the running weighted average just presented.

For the case study as developed in the previous step, applying the hanning smoother produces the sequence

4.25, 3.8125, 3.4375, 3.125, 3, 3, 3

The second entry from the left is calculated as:

$$3.8125 = \frac{1}{4}(4.25) + \frac{1}{2}(3.75) + \frac{1}{4}(3.5)$$

6. *Twice* - Smoothers based upon running medians generally tend to cause too much smoothing in a sequence and thereby remove interesting patterns. Recall that the original data set can be envisioned as being decomposed as:

   data = smooth + rough

   To recover patterns from the original series that may be still contained in the rough, one can smooth the rough sequence and then add the result to the smoothed series. Hopefully, key patterns that may have been smoothed away during the first pass of smoothing can be recovered from the rough in this manner. This operation is referred to as *reroughing*.

   For the 4253*H*, twice smoother being considered in this section, the word *twice* indicates the following calculations:

   (i) rough = data − 4253*H*,

   (ii) apply the 4253*H* smoother to the rough, and

   (iii) final smooth = 4253H (applied to given series) + 4253H (applied to rough).

   In the application the 4253*H* smooth is listed in step 5. By comparing this sequence to the original series, the rough values are found to be

   0.75, −1.8125, 0.5625, 0.875, −3, −1, 0, 1

   The first entry on the left, for example, is calculated as:

   0.75 = (5 − 4.25).

   while the third entry is:

   0.5625 = (4 − 3.4375)

   Next, the 4253*H* smoother is applied to the rough by using steps 1 to 5 with the rough data. The sequences calculated at each step are listed below:

   Step 1: Apply 4 Smoothing

   0.75, −0.53125, 0.65625, −0.625, −0.21875, −0.5, −0.5, −0.5, 1

   Step 2: Apply 2 Smoothing

0.75, 0.0625, 0.015625, –0.421875, –0.359375, –0.5, 0.1

Step 3: Apply 5 Smoothing

0.75, 0.0625, 0.015625, –0.359375, –0.359375, –0.359375, 0, 1

Step 4: Apply 3 Smoothing

0.15625, 0.0625, 0.015625, –0.359375, –0.359375, –0.359375, 0, 0.71875

Step 5: Apply Hanning

0.15625, 0.07421875, –0.06640625, –0.265625, –0.359375, –0.26953125,

–0.08984375, –0.71875

After applying the 4253*H* smooth to the rough, the last stage is to produce the final smooth by adding this to the 4253*H* smooth of the original data to get:

Final 4253*H*, twice Smooth

4.40625, 3.88671875, 3.37109375, 2.859375, 2.640625, 2.73046875, 3.08984375,

3.71875

Figures 22.3.9 and 22.3.10 display the original hypothetical series and the 4253*H*, twice smooth of the data used in the application. As can be seen, the large amount of rough contained in the given series is eliminated by using the 4253*H* twice smoother.

Electricity Consumption Application - The total annual electricity consumption for the U.S.A. is available from 1920 to 1970 in millions of kiloWatt-hours (United States Bureau of Census, 1976) and a plot of the series is displayed in Figure 4.3.10. As explained in Section 4.3.3, the most appropriate nonstationary model to fit to the square roots of this data set is an ARIMA(0,2,1) model.

The 4253*H*, twice graph of the electrical consumption series without a data transformation is shown in Figure 22.3.11. When compared to the original series depicted in Figure 4.3.10, the smooth is similar in shape to the given highly nonstationary time series. In fact, in both Figures 22.3.11 and 4.3.10, there are clearly visible trends that are increasing dramatically over time. The rough for the 4253*H*, twice smooth can be calculated using

rough = data – 4253*H*, twice.

As shown in Figure 22.3.12, there is indeed a significant rough for the series which is increasing in variance with time. In order to cause the variance to become constant or homoscedastic with time, a square root transformation is required. When comparing Figures 22.3.11 and 22.3.12, the reader should keep in mind that different multiplication factors (10^6 and 10^3, respectively) are used on the ordinate axes. Nonetheless, this example clearly illustrates that benefits can be gained by examining both the smooth and the rough for a given series.

Summary - By following steps 1 to 5, a 4253*H* smooth can be obtained for a given series. To ensure that some key characteristics of the data are not missed, in step 6 the 4253*H* smooth is applied to the rough of the smooth obtained for the original data and then the resulting smooth is added to the first 4253*H* smooth to get the 4253*H*, twice smooth. In

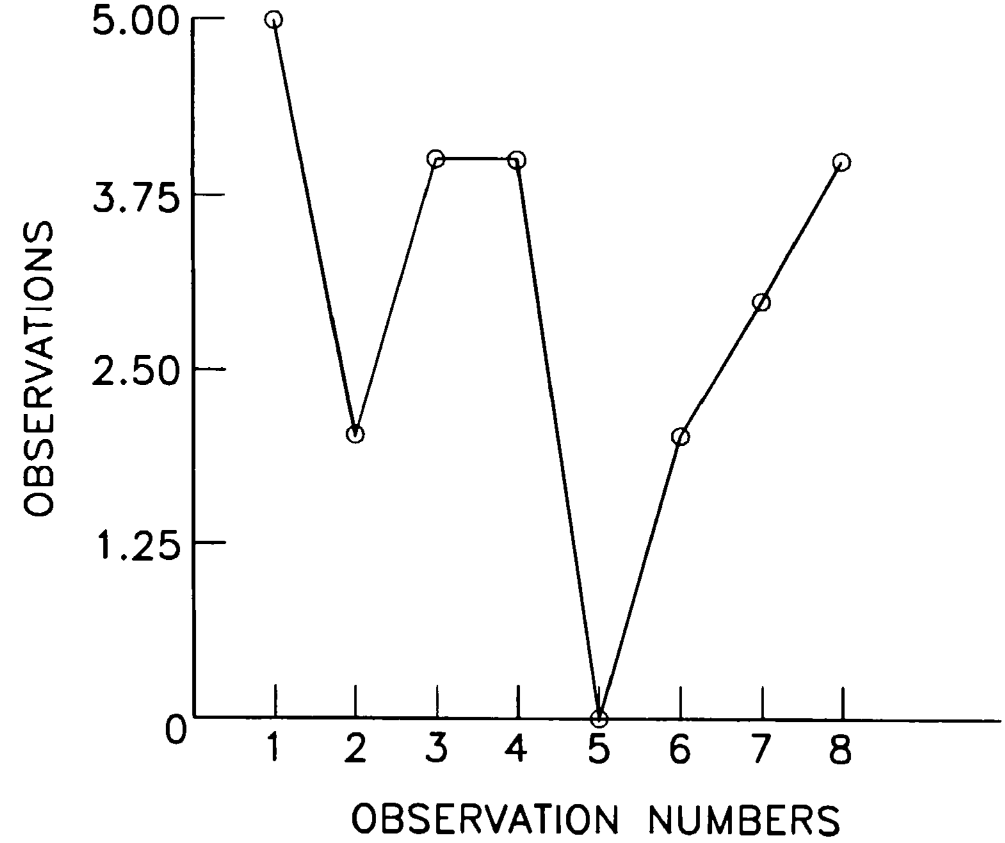


Figure 22.3.9. Graph of the original hypothetical data used in the 4253*H* smoothing example.

addition to examining plots of the 4253*H*, twice smooth to study the main statistical properties of the data, insights can also be gained by studying a plot of the rough for the 4253*H*, twice smooth.

22.3.6 Autocorrelation Function

The *ACF* at lag *k* for a given time series reflects the linear dependence between values which are separated by *k* time lags. The estimate for the ACF at lag *k* for an evenly spaced series, x_t, of length *n* can be calculated using [2.5.9] as (Jenkins and Watts, 1968)

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad k > 0 \qquad [22.3.5]$$

where $\bar{x}$ is the estimated mean of the x_t series. As noted in Section 2.5.4, the value of r_k can range from -1 to +1 where r_0 has a value of unity. Because the ACF is symmetrical about lag

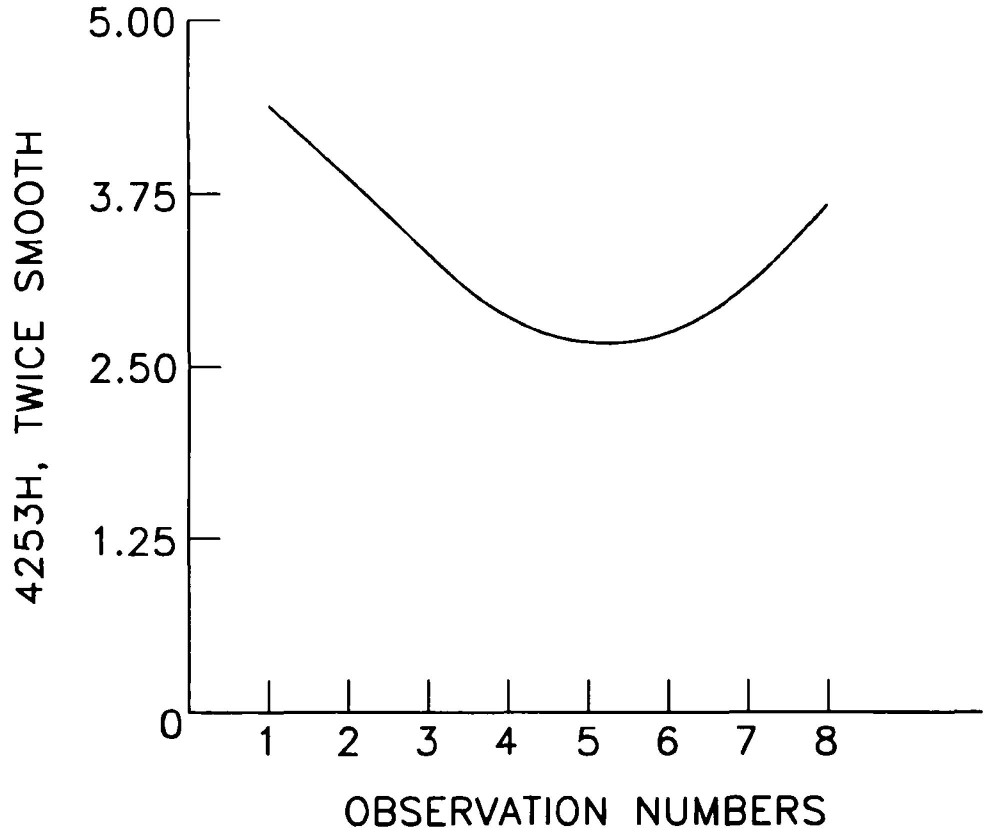


Figure 22.3.10. Plot of the 4253*H*, twice smooth for the hypothetical data of Figure 22.3.9.

zero, it is only plotted for positive lags. When the theoretical ACF is zero and, therefore, the series is white noise, r_k is asymptotically normally independently distributed with a mean of zero and variance of $1/n$. Using simulation experiments, Cox (1966) demonstrated that when r_1 is calculated for a sequence of uncorrelated samples the sampling distribution of r_1 is very stable under changes of distribution and the asymptotic normal form of the sampling distribution is a reasonable approximation even in samples as small as ten.

The ACF furnishes a method for interpreting trends in the data. If, for example, there is a large positive correlation at lag one, this means that in the plot of a series a sequence of high values will often be grouped together and low values will frequently follow other low values. In other words, when r_1 and sample ACF's at other lags are significantly different from zero, this indicates the presence of stochastic trends in the data (see Section 23.1 as well as Section 4.6 for

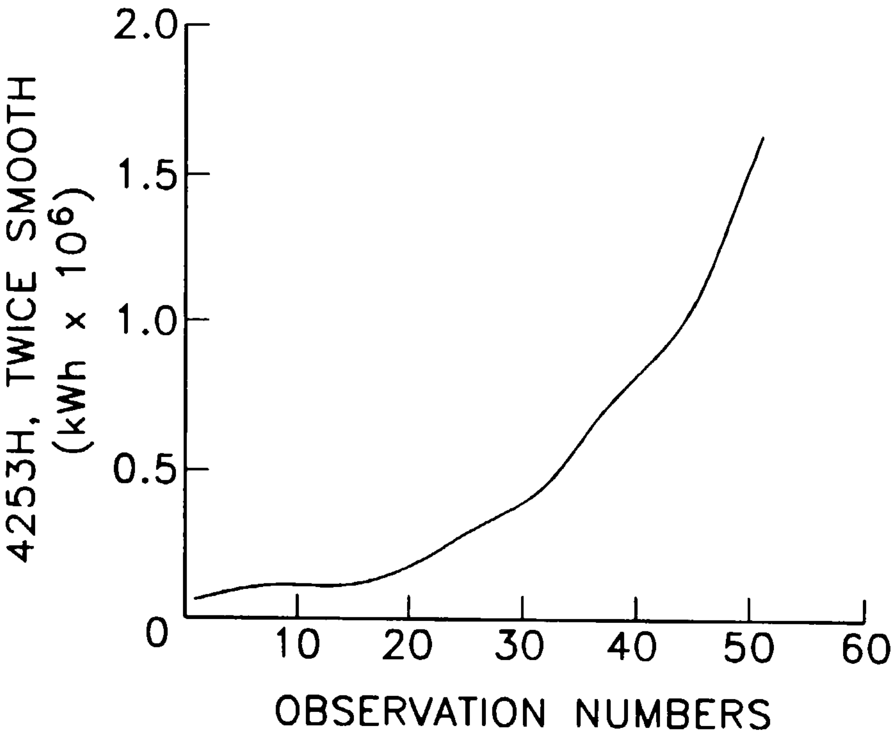


Figure 22.3.11. Plot of the 4253*H* smooth of total annual electricity consumption in the U.S.A. from 1920 to 1970.

discussions of stochastic and deterministic trends). If, for instance, the significance level is less than 0.05 this means that r_1 is significantly different from zero at the 5% significance level. The value of r_1 for the annual total organic carbon series in Figure 20.3.6 is 0.371 with a significance level of 0.137. Consequently, because the significance level of 0.137 is much larger than say the 0.05 significance level, then r_1 is not significantly different from zero. When there is an intervention which causes a significant change in the mean level of a time series such as the change shown in Figures 22.3.6 and 22.3.7 for the total organic carbon series, this introduces a trend in the data due to the observations fluctuating about different mean levels at specified sections in the series. This enforced step trend should cause a rather large value for r_1 for the entire series which is the case for the total organic carbon series. Likewise, an overall trend in the data can cause r_1 to be large. In Section 23.4, simulation experiments demonstrate that the parametric

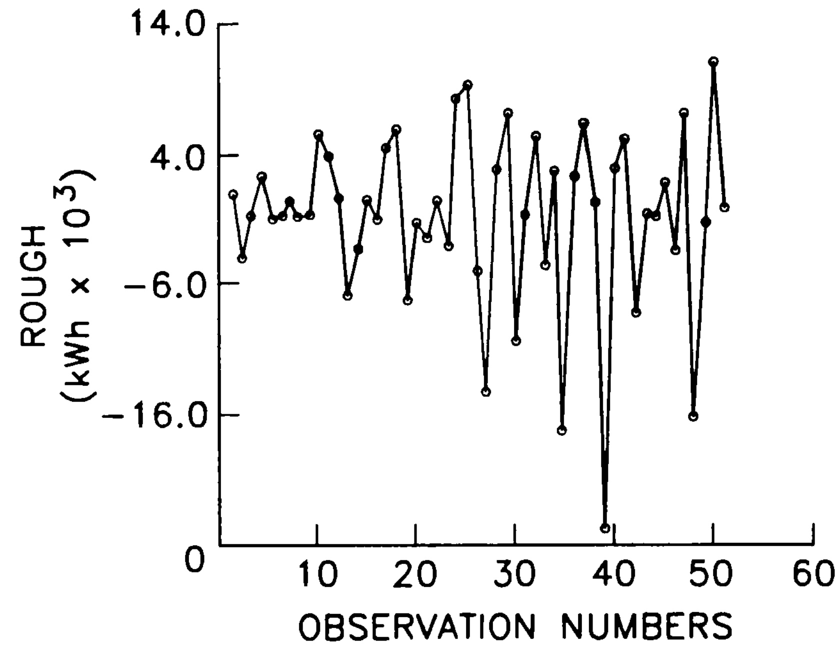


Figure 22.3.12. Rough plot for the 4253*H*, twice smooth of the total annual electricity consumption in the U.S.A. from 1920 to 1970.

test using r_1 is more powerful than the nonparametric Mann-Kendall test presented in Section 23.3.2 for detecting stochastic trends (see Sections 23.1 and 23.3.1 for a discussion and comparison of parametric and nonparametric tests). However, the Mann-Kendall test is more powerful for discovering deterministic trends.

22.4 CONFIRMATORY DATA ANALYSIS USING INTERVENTION ANALYSIS

22.4.1 Introduction

At the *exploratory data analysis* stage, important statistical characteristics of the data are discovered by employing simple graphical and numerical procedures such as those presented in Section 22.3. Figures 22.3.1 to 22.3.8 show how different exploratory data analysis tools can effectively reveal various statistical properties of the water quality time series which are studied in Section 22.3. Moreover, Figures 23.3.9 to 23.3.12 demonstrate how the 4253*H*, twice smooth and also the rough which accompanies this smooth, can uncover interesting statistical characteristics contained in a given time series. When sufficient data are available, confirmatory data analyses can be executed subsequent to the completion of the exploratory data analyses. The main objective of the *confirmatory data analysis* stage is to confirm statistically in a rigorous manner the absence or presence of certain statistical properties in the data which were uncovered by exploratory data analyses. For example, in this section the confirmatory data analysis technique of intervention analysis is used to test the hypothesis that cutting down a forest caused significant changes in the mean levels of certain water quality time series. As noted in Section 22.1, the foregoing systematic approach to data analysis is analogous to a detective solving a crime. At the exploratory data analysis stage, the sleuth collects and studies evidence which is used in rigorous court proceedings at the confirmatory data analysis stage in order to convict the suspected criminal.

Three main approaches to confirmatory data analysis are discussed in Part X of the book. The first technique is *intervention analysis* which is explained in detail in Chapter 19. As noted in Section 19.1, one of the main purposes of intervention analysis is to ascertain whether or not one or more external interventions have caused significant changes in the mean level of a time series. Another use for intervention analysis is to estimate missing observations when there are not a great number of missing values. Due to these and other uses outlined in Section 19.1, intervention analysis is extremely versatile for solving problems in an *environmental impact assessment study*. Within the water quality and quantity applications in Section 22.4.2, intervention analysis is employed for assessing the stochastic effects of an external intervention and also estimating missing values. The different forms of the intervention model used in Section 22.4.2 are described in Sections 19.3 and 19.5 of Chapter 19.

When employing intervention analysis by itself, not more than about 5% of the observations should be missing. If this is the case, intervention analysis constitutes a powerful tool for data filling and also assessing the statistical effects of interventions upon the mean level of a time series. When there are a great number of missing observations, which is the situation for the water quality time series studied in this chapter, the seasonal adjustment algorithm of Section 22.2 can be used to estimate the entries of a time series consisting of equally spaced observations. Subsequent to this, intervention analysis as well as other data analysis tools which can only be used with evenly spaced observations, can be employed.

The second and third major approaches to confirmatory data analysis are the *nonparametric tests* and the *regression analysis* methods described in Chapters 23 and 24, respectively. Although these procedures may not be as powerful as intervention analysis for checking for trends, they do not require the observations to be evenly spaced over time. Consequently, the nonparametric tests and regression analysis can be used with either evenly or unevenly spaced observations and no data filling is required.

22.4.2 Intervention Analysis Applications

Case Study

In 1961 the Marmot Creek experimental basin was established on the eastern slopes of the Rocky Mountains in Alberta, Canada (Jeffrey, 1965; Golding, 1980). The objective of the study was to determine the hydrology of the area so that guidelines which are consistent with the importance of the eastern slopes as a water supply area for Alberta and Saskatchewan, could be formulated for harvesting trees. Both the Middle Fork and Cabin Creeks are located within the Marmot basin in the province of Alberta and flows in these creeks are unregulated. Upstream from the gauging station, the area of the forested Middle Fork basin is 2.85 km^2 while the upstream area of the Cabin Creek basin is 2.12 km^2. From July to October, 1974, an intervention took place in the Cabin Creek basin when 40% of the forested area was clear-cut. Because the trees in the forested Middle Fork basin were not cut down and the basin is located close to the Cabin Creek basin, the appropriate water quality and quantity series from the Middle Fork Creek can be used as covariate series for intervention models developed for the Cabin Creek data sets. In this way the intervention components in the intervention models will more accurately measure the effects of the intervention in the Cabin Creek series.

Three different types of intervention models are developed in this section and also by McLeod et al. (1983) for solving various aspects of the problem created in the Marmot basin due to cutting down the trees in the Cabin Creek basin. The most important and interesting of the three intervention models is the third one which is called the *General Water Quality Intervention Model*. For a given water quality series for the Cabin Creek, the purpose of the third intervention model is to rigorously ascertain for which months the forest cutting intervention caused significant changes in the mean level of the water quality series. As will be seen, in order to calibrate this model, complete average monthly flow records are needed for the Cabin Creek flows which in turn are closely correlated with the Middle Fork Flows. Because there are eight missing values for the average monthly flows of Middle Fork River, an intervention model similar to the one in Section 19.3 is constructed in order to obtain efficient estimates for the missing observations. Following the development of the *Middle Fork Flow Intervention Model, The Cabin Creek Flow Intervention Model* is built for estimating four missing values in the time series of average monthly flows of the Cabin Creek and also for determining the effects of the clear-cutting intervention upon the Cabin Creek flows. In order to increase the accuracy of the Cabin Creek Flow Intervention Model, the average monthly flows of the Middle Fork River are used as a covariate series. Hence, this model is similar to the one described in detail in Section 19.5. Subsequent to the completion of the first two water quantity intervention models, a *General Water Quality Intervention Model* can be built for each of the water quality variables measured in the Cabin Creek. Because each of the water quality series consists of observations which are unevenly spaced, the data filling technique of Section 22.2 can be utilized for estimating a sequence of average monthly values. Water quality models are developed for all of the time series except the dissolved iron series, since no data are available after the intervention for this series, and also the extractable iron series. For each water quality intervention model, the covariate series are the same water quality series for the Middle Fork basin and also the monthly flows of the Cabin Creek. As an illustrative example, the procedure for fitting an intervention model to the total organic carbon series on the Cabin River, is fully explained. This water quality model constitutes an interesting version of the types of intervention models presented in Section 19.5.

Middle Fork Flow Intervention Model

Before the Middle Fork flows can be used as a covariate series in an intervention model for the Cabin Creek flows, the missing observations for the Middle Fork Creek must be estimated using a separate intervention model similar to the one in [19.3.5]. Average monthly flows are unknown for the Middle Fork Creek for April and May of 1974, February, March and April of

1975, and for January, February, and March of 1978. The entire data set for the Middle Fork flows extends from January 1964 to December 1979.

Following the notation used in Chapter 19, let y_t represent the logarithmic monthly flow of the Middle Fork Creek at time t. From [19.3.5], an intervention model for estimating the unknown observations has the form

$$y_t - \bar{y} = \sum_{j=1}^{8} \omega_{0j} \xi_{tj} + N_t \qquad [22.4.1]$$

where $\bar{y}$ is the mean of the entire y_t series, ω_{0j} is the parameter of the jth transfer function, ξ_{tj} is the jth intervention series which is assigned a value of unity where the jth observation is missing and zero elsewhere, and N_t is the noise component.

To identify the noise term, N_t, in [22.4.1], a seasonal ARIMA or SARIMA model can be fitted to the series prior to the first missing value, by utilizing appropriate model construction tools from Section 22.3. When fitting a SARIMA model to average monthly riverflows, usually it is necessary to take natural logarithms of the data and then to difference the transformed series seasonally using the operator defined in [12.2.3]. Figure 22.4.1 is a plot of the sample ACF for the seasonally differenced logarithmic data for the period from January, 1964 to December, 1973. The ACF can be calculated by substituting the data into [2.5.9] and the 95% confidence limits in Figure 22.4.1 are calculated using [12.3.1] under the assumption that the theoretical ACF is zero after lag $k = 13$. Because the ACF attenuates starting at lag 1, this may indicate the need for a nonseasonal AR parameter. Since there is a large value of the sample ACF at lag 12, this indicates that a seasonal MA parameter may be required. Consequently, following the notation from Section 12.2, it may be appropriate to fit a SARIMA $(1,0,0)\times(0,1,1)_{12}$ model from [12.2.7] to the series. The sample PACF presented in Sections 3.2.2 and 12.3.2, also confirms that this may be a reasonable model. After fitting the SARIMA $(1,0,0)\times(0,1,1)_{12}$ model to the Middle Fork data before the intervention, the ACF for the residuals can be calculated. In Figure 22.4.2, the 95% confidence limits are calculated using the formula given in [12.3.7]. Except for the residual ACF value at lag 24, all of the values lie within the 95% confidence limits and, therefore, the assumption of white noise is satisfied. The slightly large value at lag 24 could be due to chance.

Following the identification of N_t for use in [22.4.1], all of the parameters in the intervention model can be simultaneously estimated. Because natural logarithms are taken of the data, a positive quantity must be placed at the location where the observations are missing. For convenience, the logarithm of the appropriate monthly mean is substituted for each missing value in the series. In the second column of Table 22.4.1, the MLE's (maximum likelihood estimates) of the parameters and SE's (standard errors) are given for the eight ω_{0j} intervention parameters in [22.4.1]. From Section 9.3.3, the estimate of the missing observation in the logarithmic or transformed domain is

$$\hat{y}_{t_j} = \ln \textit{monthly mean} + \hat{\omega}_{0j}$$

where t_j is the time for which the observation at t_j is missing. By taking the inverse logarithmic transformation, the estimate of the missing observation in the untransformed domain is

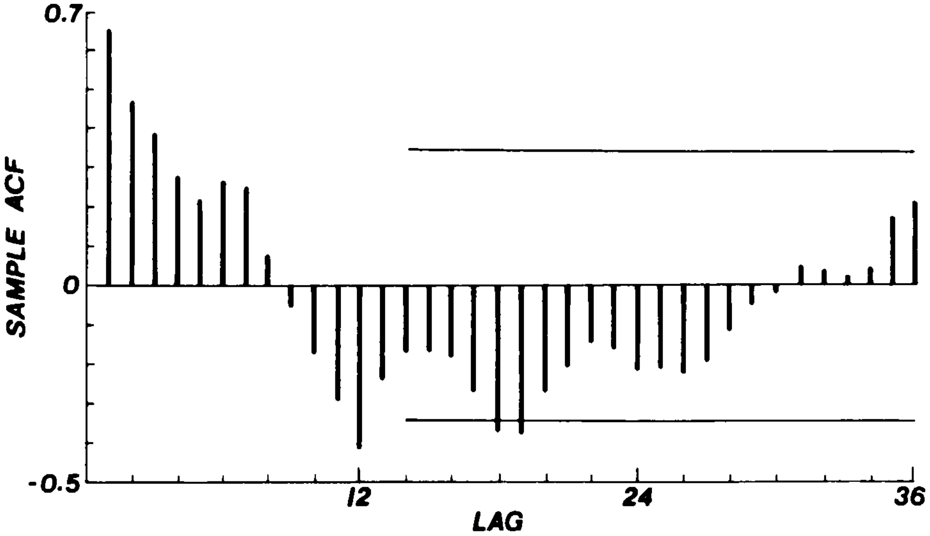


Figure 22.4.1. Sample ACF for the seasonally differenced logarithmic Middle Fork flows before the intervention.

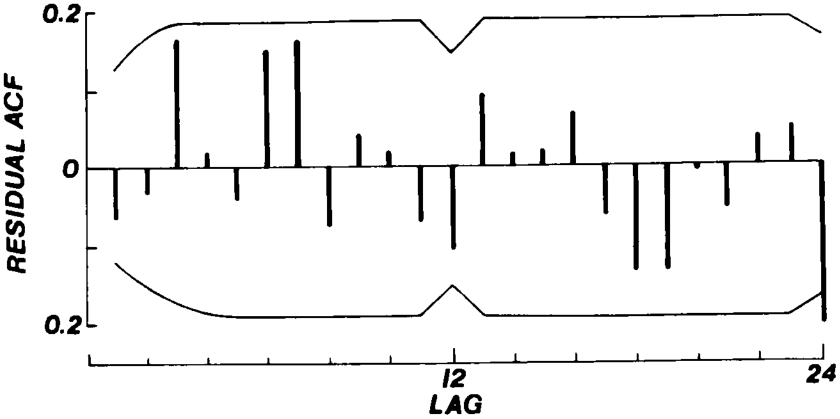


Figure 22.4.2. Residual ACF for the Middle Fork SARIMA model for before the intervention.

$$\hat{Y}_{t_j} = e^{\hat{y}_{t_j}} = \exp(\ln \textit{monthly mean} + \hat{\omega}_{0j}) \qquad [22.4.2]$$

The estimates for the missing monthly observations in m^3/s are displayed in the last column in Table 22.4.1.

Table 22.4.1. Estimated parameters for the Middle Fork Flow intervention model.

Parameter	Estimate of ω_{0j} ± Standard Error	Date of Missing Value	Estimate of Missing Value in m^3/s
ω_{01}	0.6029 ± 0.2976	April, 1974	0.0219
ω_{02}	-0.5316 ± 0.2959	May, 1974	0.0348
ω_{03}	0.1849 ± 0.3083	Feb., 1975	0.0048
ω_{04}	0.3057 ± 0.3456	March, 1975	0.0041
ω_{05}	0.6885 ± 0.3084	April, 1975	0.0139
ω_{06}	-0.1244 ± 0.3150	Jan., 1978	0.0044
ω_{07}	-0.0236 ± 0.3571	Feb., 1978	0.0039
ω_{08}	-0.0838 ± 0.3164	March, 1978	0.0028

Cabin Creek Flow Intervention Model

Because the missing values in the Middle Fork riverflow series have all been estimated, the complete data set can now be used as a covariate series for an intervention model for the Cabin Creek. However, there are four missing observations for the Cabin Creek which occur in February, March, April, and May of 1979. Consequently, in addition to the clear-cutting intervention which took place from July to October of 1974, intervention components must be included in the model so that the four missing values can be estimated.

Some of the exploratory data analysis tools from Section 22.3 can be employed to check if there appear to be changes in the Cabin Creek flows due to the forest cutting intervention. For example, box-and-whisker graphs from Section 22.3.3 can be constructed for the series both before and after the intervention date. When the medians for each month before and after the intervention are compared, the median levels do not appear to change very much. Likewise, when other exploratory tools are employed, no noticeable changes in the flow series are detected as a result of cutting down the forest.

To ascertain rigorously if clear-cutting the forest has significantly affected the mean levels of the average monthly flows of the Cabin Creek, an appropriate intervention model can be constructed at the confirmatory data analysis stage by following the three steps of identification, estimation and diagnostic checking described in Section 19.5.3. Following the general format of the model in [19.5.8], the intervention model for the flows of the Cabin Creek can be written as

$$y_t - \bar{y} = \sum_{i=1}^{12} \omega_{0i}\xi_{ti} + \sum_{j=13}^{16} \omega_{0j}\xi_{tj} + \omega_{017}(x_t - \bar{x}) + N_t \qquad [22.4.3]$$

where y_t is the monthly logarithmic Cabin Creek flows and $\bar{y}$ is the mean of the entire y_t series, ω_{0i} is the transfer function parameter for the ith monthly intervention where there is one parameter for each month of the year, ξ_{ti} is a monthly step intervention which is given a value of unity for the month it represents during and after the intervention but given a value of zero elsewhere, ω_{0j} is the transfer function parameter for a missing data point where there are four such parameters, ξ_{tj} is the intervention series for ω_{0j} where it is given a value of unity at the time that the observation is missing and a value of zero elsewhere, x_t is the logarithmic Middle Fork flows and $\bar{x}$ is the mean of the entire x_t series, ω_{017} is the transfer function parameter for the covariate x_t series, and N_t is the correlated noise term.

The transfer functions in [22.4.3] are designed from a physical understanding of the problem. Since it would be expected that the Middle Fork and Cabin Creek flows would be closely related during the same month due to common climatic conditions, the parameter ω_{017} is included as the parameter in the covariate transfer function. Because only six observations for each month are available after the intervention, the step intervention series, ξ_{ti}, along with the ω_{0i} parameter is included in the first term for each month on the right hand side in [22.4.3]. When more data becomes available, it may be reasonable to include a parameter in the denominator of each transfer function that models the clear-cutting intervention. In [19.5.10], it is shown how a term in the denominator can model the attenuating affects of a forest fire upon riverflows as the forest slowly recovers over the years. In this study, transfer functions of the form $\dfrac{\omega_{0i}}{1 - \delta_{li}B^{12}}$ were included in the first summation term on the right hand side of [22.4.3] but meaningful results were not obtained due to the lack of sufficient data and a long enough time period after the intervention. Perhaps after about ten years, enough data will be available so that two parameters can be included and thereby allow the impacts of forest recovery to be more fully explored within the structure of the intervention model.

After designing the transfer functions in [22.4.3], the noise term, N_t, must be identified. As noted throughout Chapter 19, a convenient procedure to employ is to first assume that N_t is white noise. The parameters and residuals in [22.4.3] can then be estimated. Since the residuals will probably not be white noise, a SARIMA model can be identified for fitting to the residuals. For the case of the Cabin Creek residuals, the most appropriate model is found to be a SARIMA $(1,0,0)\times(1,0,0)_{12}$ model. Assuming this form for N_t, the parameters for the complete model in [22.4.3] are simultaneously estimated again. Diagnostic checks applied to the residuals from the latest model design, demonstrate that the model is satisfactory since the residuals are uncorrelated. In practice, various forms of the transfer functions and noise term must be tried before a satisfactory model is found. Consequently, the model building procedure is not quite as simple as it may appear in the foregoing explanation. Experience, coupled with a sound understanding of both the physical problem and the capabilities of the intervention model, help to reduce the time required to design an appropriate intervention model.

The parameter estimates and SE's for the four missing observations are given in Table 22.4.2. Since natural logarithms are taken of the data, it is necessary to include positive values at the four locations where the observations are missing. The logarithm of the appropriate average monthly value across all the years is substituted at each location where an observation is missing. After calibrating the complete intervention model written in [22.4.3], each estimated missing value can be calculated using [22.4.2]. In the last column in Table 22.4.2, the estimated monthly values ore displayed.

Table 22.4.2. Estimated missing values for the Cabin Creek flows.

Parameter	Estimate ± Standard Error	Date of Missing Value	Estimate of Missing Value
ω_{013}	0.2893 ± 0.2300	Feb., 1979	0.004
ω_{014}	0.1351 ± 0.2607	March, 1979	0.003
ω_{015}	0.6682 ± 0.2611	April, 1979	0.012
ω_{016}	0.5834 ± 0.2300	May, 1979	0.079

The MLE for ω_{017}, which is the transfer function parameter for the covariate series consisting of the logarithmic Middle Fork flows, is 0.814 with a SE of 0.020. Because $\hat{\omega}_{017}$ is much larger than 1.96 times the SE, it is significantly different from zero. Accordingly, it is worthwhile to include the covariate series in the intervention model in [22.4.3] in order to enhance the credibility and accuracy of the model.

The MLE's and SE's for the twelve intervention parameters for the clear-cutting, are presented in Table 22.4.3. To calculate the percentage change in the mean level of a specific monthly flow due to the intervention, the following formula given in [19.2.20] is employed.

$$\% \, change = (e^{\hat{\omega}_{0i}} - 1)100 \qquad [22.4.4]$$

To calculate the 95% confidence limits simply add and subtract 1.96 times the SE to the estimated ω_{0i} and then substitute these two values into [22.4.4]. The percentage change in the mean level for each month along with the 95% confidence levels are presented in Table 22.4.3. For nine out of twelve months, zero falls within the 95% confidence limits. Therefore, for these months it can be argued that the percentage changes in the mean levels are not significantly different from zero. However, for January, March and November, zero does not fall within the 95% confidence limits and, consequently, there appear to be significant changes in the mean levels for these months. For January and March the mean levels have decreased while for November the average flow has risen. Nevertheless, notice that for each of these three months, one side of the limits for the 95% confidence limits, is quite close to zero. Consequently, to simplify the intervention model developed in the next subsection, it is assumed that the clear cutting of the forest has not significantly altered the Cabin Creek flows.

Table 22.4.3. Estimated parameters for modelling the intervention effects in the Cabin Creek flow intervention model.

Month	Parameter	Estimate ± Standard Error	Percentage Change	95% Confidence Interval
Jan.	ω_{01}	-0.2303 ± 0.1153	-20.57	-36.64, -0.43
Feb.	ω_{02}	-0.1930 ± 0.1183	-17.55	-34.62, 3.97
March	ω_{03}	-0.2353 ± 0.1191	-20.96	-37.42, -0.18
April	ω_{04}	-0.1652 ± 0.1192	-15.23	-32.88, 7.07
May	ω_{05}	-0.0931 ± 0.1197	-8.89	-27.94, 15.20
June	ω_{06}	-0.0682 ± 0.1206	-6.59	-26.25, 18.31
July	ω_{07}	-0.0872 ± 0.1290	-8.35	-28.83, 18.01
Aug.	ω_{08}	-0.1353 ± 0.1411	-12.65	-33.75, 15.17
Sep.	ω_{09}	0.0802 ± 0.1454	8.35	-18.51, 44.08
Oct.	ω_{010}	-0.1468 ± 0.1419	-13.65	-34.62, 14.04
Nov.	ω_{011}	0.3007 ± 0.1396	35.08	2.75, 77.58
Dec.	ω_{012}	-0.0337 ± 0.1289	-3.31	-24.90, 24.49

General Water Quality Intervention Model

Intervention models were developed for twelve water quality variables on the Cabin Creek although representative results are only shown in this section for the total organic carbon intervention model. For each water quality intervention model, the covariate series are the same water quality series for the Middle Fork basin and also the monthly flows of the Cabin Creek. Qualitatively, the General Water Quality Intervention Model is written as

$$\begin{array}{ccccccccc} \textit{Cabin} & & & & & & \textit{Middle} & & \\ \textit{Creek} & & \textit{monthly} & & \textit{Cabin} & & \textit{Fork} & & \\ \textit{water} & = & \textit{interventions} & + & \textit{Creek} & + & \textit{water} & + & \textit{Noise} \\ \textit{quality} & & & & \textit{flows} & & \textit{quality} & & \\ \textit{series} & & & & & & \textit{series} & & \end{array}$$

Mathematically, appropriate components from the finite difference equation in [19.5.2] can be utilized in order to write the General Water Quality Intervention Model as

$$y_t - \bar{y} = \sum_{i=1}^{12} \omega_{0i}\xi_{ti} + \omega_{013}(x_{t1} - \bar{x}_1) + \omega_{014}(x_{t2} - \bar{x}_2) + N_t \qquad [22.4.5]$$

where y_t is the average monthly water quality series for the Cabin Creek that was estimated using the seasonal adjustment algorithm of Section 22.2, $\bar{y}$ is the mean of the y_t series, ξ_{ti} is the intervention series for a given month where it is assigned a value of one for the month it represents from the intervention onwards and a value of zero elsewhere, ω_{0i} is the transfer function parameter for the ξ_{ti} series and the MLE for ω_{0i} can be used to ascertain the effects of the intervention for the month being studied, x_{t1} is the estimated monthly logarithmic series for the

Cabin Creek where the seasonal adjustment algorithm in Section 22.2 is used to estimate the monthly flows from daily flows that occur at the same time as the water quality observations, $\bar{x}_1$ is the mean of the x_{t1} series, ω_{013} is the transfer function for the Cabin Creek flow series, x_{t2} is the same estimated monthly water quality series as y_t but for the Middle Fork Creek and the seasonal adjustment algorithm is used to estimate x_{t2}, $\bar{x}_2$ is the mean of the x_{t2} series, ω_{014} is the transfer function parameter for the covariate Middle Fork water quality series, and N_t is the noise term which can be modelled by an appropriate SARMA or SARIMA model from Chapter 12.

In [22.4.5] the seasonally adjusted monthly flows are employed as a covariate series, x_{ti}. The reason for using the seasonally adjusted series rather than the known monthly riverflows is that this may help to eliminate any problems due to seasonal adjustment that are contained in the y_t series. It should be kept in mind that by considering the flows as a covariate series, the stochastic or statistical relationship between the flow, x_{t1}, and the water quality series, y_t, is formally modelled through the transfer function parameter, ω_{013}, in the overall intervention model in [22.4.5].

When constructing the water quality intervention models in [22.4.5], the identification, estimation and diagnostic check stages of model development described in Section 19.5.3 are adhered to. Although the transfer functions for all the water quality series are the same as those in [22.4.5], it should be pointed out that quite a few different types of transfer functions were actually tested. For instance, because not too many observations for each month are available after the intervention, a step intervention along with a ω_{0i} parameter is included in the first term for each month on the right hand side in [22.4.5]. As also noted for the intervention model in [22.4.3], if more data were available the possibility of including a parameter in the denominator of each transfer function would have been feasible. In [19.5.10] within Section 19.5.4, it is explained how a term in the denominator can model the attenuating effects of a forest fire upon riverflows as the forest slowly recovers over the years. Finally, a specific SARIMA model had to be identified separately for modelling N_t in [22.4.5] for each water quality intervention model. The same procedure described for the Cabin Creek flow intervention model was employed to design a specific noise term for each intervention model.

Total Organic Carbon Application: As shown by the applications in Section 22.3, exploratory data analyses clearly detect the effects of the forest clearing upon the total organic carbon series for the Cabin Creek. For example, when the box-and-whisker graphs for before and after the intervention are compared in Figures 22.3.4 and 22.3.5, respectively, the decrease in the median level after the intervention can be easily seen for almost all the months. Likewise, the average annual plot in Figure 22.3.6 and the blurred smooth in Figure 22.3.7 clearly detect the drop in the mean level of total organic carbon in later years. Finally, since the value of the ACF at lag one calculated using [22.3.5] for the annual series is significantly different from zero, this suggests the presence of a trend in the data.

The foregoing exploratory facts are rigorously confirmed in a statistical sense by fitting the intervention model in [22.4.5] to the total organic carbon series which is available from the start of 1971 to the end of 1978. Natural logarithms are used for the two total organic carbon series given by y_t and x_{t2} for the Cabin and Middle Fork Creeks, respectively. The SARIMA model identified for the noise term, N_t, contains one nonseasonal AR parameter and one seasonal AR parameter, and as explained in Section 12.2.2, it can be written as $(1,0,0)\times(1,0,0)_{12}$. The parameter, ω_{013}, which relates the Cabin Creek flows to the total organic carbon in the Cabin Creek has a MLE of 0.081 with a SE of 0.095. Since the MLE of ω_{013} is about the same size as its SE, it may be worthwhile to include the flows as a covariate series in the intervention model. The MLE for ω_{014} is 0.620 with a SE of 0.082 and, consequently, it is very informative to incorporate the covariate total organic series from Middle Fork Creek into the model. In Table 23.4.4, the MLE's and SE's are presented for the twelve intervention parameters contained in the first component on the right hand side of [22.4.5]. Also included in Table 23.4.4 is the percentage change in mean level for each month along with the 95% confidence limits which are calculated using [22.4.4]. For all the months where zero is not included in the 95% confidence limits, the percentage change in the mean level is confirmed to be significantly different from zero. Accordingly, from Table 22.4.4 it can be seen that there is a significant drop in the mean level of total organic carbon in the Cabin Creek during the summer months of June, July and August.

Table 22.4.4. Intervention parameter estimates for the total organic carbon intervention model for the Cabin Creek.

Month	Parameter	MLE	Standard Error	Percentage Change	95% Confidence Limits	
January	ω_{01}	0.002	0.231	0.17	-36.33,	57.59
February	ω_{02}	0.085	0.231	8.92	-30.71,	71.22
March	ω_{03}	-0.169	0.227	-15.52	-45.86,	31.82
April	ω_{04}	-0.216	0.224	-19.42	-48.11,	25.11
May	ω_{05}	-0.053	0.228	-5.12	-39.25,	48.20
June	ω_{06}	-0.716	0.227	-51.12	-68.69,	-23.68
July	ω_{07}	-0.524	0.256	-40.81	-64.18,	-2.20
August	ω_{08}	-0.566	0.260	-43.21	-65.88,	-5.49
September	ω_{09}	0.026	0.260	2.65	-38.34,	70.90
October	ω_{010}	-0.019	0.258	-1.91	-40.86,	62.69
November	ω_{011}	0.048	0.266	4.90	-37.75,	76.79
December	ω_{012}	-0.344	0.270	-29.13	-58.26,	20.32

22.5 CONCLUSIONS

To execute a comprehensive data analysis study, one can follow the exploratory data analysis and confirmatory data analysis stages. As demonstrated by water quality applications, this approach is especially effective for detecting and modelling trends which may be contained in messy environmental data. The time series being analyzed may be very messy because the series may possess various handicaps such as having missing observations, being nonnormally distributed, possessing outliers and being short in length. Nevertheless, by employing appropriate exploratory and confirmatory data analysis tools, as much useful information as possible can be gleaned from the available data, even if the quality and quantity of the data are not very good.

The purpose of the exploratory data analysis stage is to uncover important statistical properties of the data by utilizing simple graphical and numerical tools, some of which are discussed in detail in Section 22.3. Some exploratory techniques such as a graph of the series against time (Section 22.3.2), box-and-whisker graphs (Section 22.3.3) and the cross-correlation function (Section 22.3.4), do not require that the time series be evenly spaced over time. On the other hand, other exploratory data analysis techniques like Tukey smoothing (Section 22.3.5) and the ACF (Section 22.3.6) are designed to be used with data points that are equally spaced over time. Fortunately, a number of flexible data filling procedures are now available for estimating the entries of an evenly spaced time series from a data set for which the time intervals between adjacent observations are not the same. Depending upon how much information is missing and the number of observations available, an appropriate data filling technique can be selected from Section 22.2, 19.3, or 18.5.2. Subsequent to filling in missing observations, one can employ suitable exploratory and confirmatory data analysis tools which require evenly spaced measurements.

At the confirmatory data analysis stage, three different types of approaches which can be used to rigorously characterize trends are intervention analysis, nonparametric tests and regression analysis, described in Section 22.4 and Chapter 19, Chapter 23, and Chapter 24, respectively. Nonparametric tests and regression analysis can be used with unequally or equally spaced data whereas intervention analysis must be employed with an evenly spaced sequence of observations. As demonstrated in this chapter and also Chapter 19, the intervention model constitutes an extremely powerful and comprehensive parametric model which can accurately model the magnitude and shape of a trend caused by a known intervention. Furthermore, as explained in Section 22.4.2, the impacts of water quantity upon water quality can be realistically incorporated into the intervention model by including the water quantity time series as a covariate series in the intervention model in [22.4.5].

In many situations, an analyst is requested to execute a comprehensive data analysis study in order to discover and model trends after the data have already been collected by other people. As a result, the data may be rather messy and thereby difficult to model. Of course, there are no data analysis tools that can extract information which is not contained in the data to begin with. Nonetheless, by using the most suitable data analysis techniques, the maximum amount of useful information can be discovered and modelled. When the analyst can assist in optimally designing the data collection scheme, then some of the suggestions given in Section 19.7 and elsewhere may be helpful.

PROBLEMS

22.1 The seasonal adjustment algorithm in Section 22.2 is described for estimating average monthly values for a time series using daily values that are available at irregular time intervals. Explain how this algorithm would work for the following situations:

(a) estimating average quarterly values, and

(b) estimating weekly values.

22.2 Select a daily time series for which all of the observations are available over a ten year time period. Randomly remove about 70% of the daily data and then employ the seasonal adjustment algorithm of Section 22.2 to estimate the average monthly values. Compare these monthly estimates to those obtained when the complete set of daily values are employed for determining the average monthly values.

22.3 In Section 22.3, some useful exploratory data analysis tools are presented. By referring to an appropriate reference on exploratory data analysis, describe three other exploratory data analysis techniques which are not discussed in Section 22.3. You may, for instance, wish to write about stem-and-leaf displays. Be sure to mention the main statistical characteristics that each method is designed to uncover in a given data set.

22.4 Select an average monthly time series that is of interest to you. Obtain a plot of the observations over time as well as a box-and-whisker graph for each season. Describe the main statistical characteristics contained in your data set which are graphically revealed using each exploratory data analysis technique. Which of the two graphical methods was most helpful for better understanding the statistical and stochastic properties of your data?

22.5 As mentioned in Section 22.3.3, a seasonal notched box-and-whisker graph can be employed for graphically testing whether medians across two or more seasons are significantly different. Explain why using seasonal notched box-and-whisker plots in this way is equivalent to graphically carrying out a formal hypothesis test (see Section 23.2.2 for a review of hypothesis tests). In your explanation, be sure to clearly state the null and alternative hypotheses as well as the test statistic. Assuming normality and independence of the data for each season, derive the 5% significance level for the test. Finally, state advantages of graphically implementing statistical tests as part of exploratory data analysis tools.

22.6 Select a set of water quality time series measurements that are available for a variety of water quality variables measured in a river or lake. Following the procedure of Section 22.3.4, determine the cross-correlation matrix for these series. When commenting upon your results use physical explanations of the phenomena to help confirm what is found statistically.

22.7 Choose an annual time series which you suspect may contain trends. For this series, plot the following graphs and then comment upon your findings regarding the main statistical characteristics of the series. Be sure to make comparisons across the graphs and clearly point out the advantages of studying each type of graphical output.

(a) Plot of the data,

(b) Box-and-whisker graph,

(c) Blurred 3RSR smooth,

(d) 4253H, twice smooth,

(e) Rough for the 4253H, twice smooth.

22.8 Carry out the instructions of 22.7 for a seasonal time series of your choice.

22.9 Select an annual time series to which you apply all of the most appropriate exploratory data analysis tools of Section 23.3 and elsewhere. Justify the reasons for choosing these exploratory techniques and summarize your main statistical findings.

22.10 Execute the instructions of problem 22.9 for a seasonal time series.

22.11 By referring to an appropriate reference, find a smoother not covered in Section 22.3.5 which you think may work well in practice. Outline the steps that are followed when applying this smoother to a time series. Assess the main capabilities and weaknesses of the smoother. Finally, apply this smoother to a time series of your choice and comment upon your statistical findings.

22.12 In a column on the left hand side of a page, write down a fairly extensive list of statistical characteristics which you would like exploratory data analysis tools to discover when examining water quality time series. In a row across the top of the page copy down the names of a variety of informative exploratory data analysis techniques. Then, below each technique put check marks opposite the statistical properties that the method is designed to find when these properties are present in the data. Explain how this table that summarizes the capabilities of exploratory data methods can be useful in a case study.

22.13 Carry out the instructions of problem 22.12 for hydrological time series.

22.14 Execute the instructions of problem for 22.12 for meteorological data sets.

22.15 Select a set of water quantity and quality time series that may have been significantly influenced by a known external intervention.. Carry out a comprehensive data analysis of these time series by employing appropriate exploratory and confirmatory data analysis tools in order to detect and model possible trends as well as other interesting statistical characteristics.

REFERENCES

DATA SET

United States Bureau of the Census (1976). *The Statistical History of the United States from Colonial Times to the Present.* United States Government.

EXPERIMENTAL BASINS

Golding, D. L. (1980). Calibration methods for detecting changes in streamflow quantity and regime. In *The Influence of Man on the Hydrological Regime with Special Reference to Representative and Experimental Basins, Proceedings of the Helsinki Symposium,* IAHS (International Association of Hydrological Sciences), 130:3-7.

Jeffrey, W. W. (1965). Experimental water sheds in the Rocky Mountains, Alberta, Canada. In *Proceedings of the Symposium on Representative and Experimental Areas, Budapest Symposium*, 6:502-521 IAHS (International Association of Hydrological Sciences) 66.

EXPLORATORY DATA ANALYSIS

Barnett, V. (1975). Probability plotting methods and order statistics. *Applied Statistics* 24(1):95-108.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1981). Wastewater treatment: A review of statistical applications. In *Environmetrics 81: Selected Papers, Selections from USEPA-SIAM-SIMS Conference,* Alexandria, Virginia, pages 77-99.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* 74(368):829-836.

Cluis, D. A. (1983). Visual techniques for the detection of water quality trends: Double-mass curves and CUSUM functions. *Environmental Monitoring and Assessment,* 3:173-184.

Cox, D. R. (1966). The null distribution of the first serial correlation coefficient. *Biometrika,* 53:623-626.

Cunnane, C. (1978). Unbiased plotting positions - a review. *Journal of Hydrology,* 37:205-222.

du Toit, S. H. C., Steyn, A. G. W., and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis*. Springer-Verlag, New York.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association,* 71(354):378-385.

Hewlett-Packard (1977). *HP-29C Applications Book.*

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, New York.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications.* Holden-Day, San Francisco.

Mallows, C. L. (1980). Resistant smoothing. In Anderson, O. D., editor, *Time Series, Proceedings of the International Conference held at Nottingham University,* March 1979, pages 147-155. North-Holland.

McGill, R., Tukey, J. W. and Laren, W. A. (1978). Variation of Box plots. *The American Statistician,* 32(1):12-16.

McNeil, D. R. (1977). *Interactive Data Analysis.* Wiley, New York.

Ramsey, F. L. (1988). The slug trace. *The American Statistician,* 42(4):290.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts.

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association,* 75:609-615.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston.

SEASONAL ADJUSTMENT

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3-32.

Granger, C. W. J. (1980). *Forecasting in Business and Economics*. Academic Press, New York.

Hillmer, S. C. and Tiao, G. C. (1985). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association*, 77:63-70.

Kendall, M. G. (1973). *Time-Series*. Hafner Press, New York.

Shiskin, J., Young, A. H. and Musgrave, J. C. (1976). The x-11 variant of the census method II seasonal adjustment program. Technical report number BEA-76-01, Bureau of Economic Analysis, U.S. Department of Commerce, Washington, D.C.

TREND ASSESSMENT STUDIES

McLeod, A. I., Hipel, K. W. and Camacho, F. (1983). Trend assessment of water quality time series. *Water Resources Bulletin*, 19(4):537-547.

CHAPTER 23

NONPARAMETRIC TESTS FOR TREND DETECTION

23.1 INTRODUCTION

As demonstrated by the water quality and quantity applications in Sections 22.4 and 19.2 to 19.5, *intervention analysis* constitutes a flexible tool for rigorously ascertaining the effects of interventions upon the mean level of a series. Because the intervention model in [19.5.8] contains parameters which can be conveniently estimated when the innovations are assumed to be normally independently distributed, this model is referred to as a *parametric model.* Even when the data are quite messy and contain many missing values, after employing a technique such as the seasonal adjustment procedure of Section 22.2 to fill in the missing observations, the water quality applications of Section 22.4.2 demonstrate that an appropriate intervention model can then be conveniently calibrated to the estimated series of evenly spaced data. Diagnostic checks of the model residuals of the fitted water quality intervention model in [22.4.5], as well as all the other intervention models fitted to time series in Chapters 19 and 22, confirm that the underlying assumptions of intervention models can be readily satisfied in practical applications. Consequently, the intervention model is a powerful parametric statistical technique for use in *environmental impact assessment.*

Another family of parametric models for employment in environmental impact assessment is the *regression analysis* set of models described in Chapter 24. An advantage of regression analysis is that it can be used with both evenly or unevenly spaced observations.

In order to lessen the number of underlying assumptions required for testing a hypothesis, such as the presence of a specific kind of trend in a data set, researchers developed *nonparametric tests.* Because a nonparametric test is a method for testing a hypothesis whereby the test does not depend upon the form of the underlying distribution of the null hypothesis, a nonparametric test is often referred to as a *distribution free or distribution independent method.* As a matter of fact, some distribution free methods assume there are parameters in the models which form the basis for the tests whereas other distribution free tests do not involve any parameters, either directly or indirectly in the tests. Although the term nonparametric should be confined to describing distribution free tests for which there are no parameters, in practice it has been interpreted as standing for the set of all distribution free methods. Hence, within this text the more commonly used phrase of nonparametric tests will be used even though it is more correct to utilize the expression of distribution free tests.

Nonparametric tests were developed for use in environmental impact assessment because scientists were concerned that the statistical characteristics of *messy environmental data* would make it difficult to use parametric procedures. As noted by Hirsch and Slack (1984), natural time series may contain one or more of a number of properties which are undesirable for use with parametric tests. In particular, hydrologic and water quality data may be nonnormally distributed and follow a distribution which is usually positively skewed. Because the adoption of proper sampling procedures are often not considered, environmental time series are not commonly measured at uniform time intervals. Moreover, data are often *censored* by only listing measurements below a certain level as being "less than" or measurements above a specified level as being "greater than". For instance, concentration values for metals or organic compounds which fall below the *detection limits* for certain chemical tests are reported simply as less than the limits of detection. Fortunately, the foregoing and other characteristics of environmental time series can often be properly accounted for in order to make the data suitable for use with parametric testing. For example, as noted at many locations in this text, invoking a data transformation, such as the Box-Cox transformation in [3.4.30], can often help to alleviate nonnormality, although this is not always the case. Depending upon how much data are missing, an appropriate *data filling* procedure can be selected from Sections 22.2, 19.3, or 18.5.2 to obtain an estimated evenly spaced time series. When observations are given as less than the detection limit, one way to estimate what they should be is to consider them to be missing and estimate them using a suitable data filling procedure.

Nonparametric tests have few underlying assumptions and tend to ignore the magnitude of the observations in favour of the relative values or *ranks* of the data. As a result, a given nonparametric test which is designed, for instance, for checking for the presence of a trend, may only provide a yes or no answer as to whether or not a trend may be contained in the data. The output from the nonparametric test may not give an indication of the type or magnitude of the trend. In order to have a more precise test about what is occurring, many assumptions must be made and as more and more assumptions are formulated, nonparametric tests begin to look more and more like parametric tests. As a matter of fact, as noted by Savage in the encyclopedia edited by Kruskal and Tanur (1978, p. 637), the dividing line between nonparametric and parametric tests is not a sharp one. Finally, Conover and Iman (1981) explain how rank transformations of data sets act as a bridge between parametric and nonparametric statistics.

Because nonparametric tests are usually designed to indicate the presence but not the magnitude of a given statistical characteristic, some authors consider them to be exploratory data analysis procedures. Nonetheless, as is explained for the nonparametric tests described in Section 23.3, and Appendices A23.1 to A23.3, all these tests are designed for specifically testing certain hypotheses. Since they are utilized for *hypothesis testing*, within this text nonparametric tests are deemed to be confirmatory data analysis tools. Of course, after detecting the presence of a trend using a nonparametric test, a more powerful confirmatory data analysis method such as the intervention analysis technique of Section 22.4 and Chapter 19, can be employed for obtaining precise statistical statements about the trends.

Over the years practitioners have argued about whether nonparametric or parametric tests should be employed. An *advantage* of nonparametric tests is that they are distribution free and hence fewer assumptions have to be made about the data. On the other hand, as shown for the intervention model in Section 22.4, often many difficulties with the data which appear to make the series unusable with a parametric technique, can, in fact, be overcome. Cox and Hinkley (1974, Section 6.1) describe a number of *drawbacks* to nonparametric tests which they say limit their practical importance. One of the limitations is that when a parametric test is appropriate, a nonparametric test cannot be as powerful as the most efficient parametric test. Additionally, the results from a nonparametric test often do not adequately describe what is happening with a data set. In order to achieve a reasonable description and understanding of the system under investigation in concise and simple terms, a parsimonious parametric model is required where each parameter describes some important aspect of the system. Keeping in mind the assets and limitations of both parametric and nonparametric tests, a pragmatic approach to data analysis may be to use whatever tests seem to be most appropriate, whether they are nonparametric or parametric tests. For example, when analyzing vast amounts of environmental data for the presence of trends, in conjunction with the exploratory data analysis tools of Section 22.3, nonparametric testing can be used to locate the data which contain trends. Parametric techniques for detecting trends in a time series are referenced in Sections 19.2.3 and 24.2.1. Subsequent to a perusal of the written historical records to find physical causes for trends in the data, intervention analysis (Chapter 19 and Section 22.4) or regression analysis (Chapter 24) can be employed for obtaining rigorous statistical statements about the types and magnitudes of the trends. Bloomfield et al. (1983), Bloomfield (1992) and Bloomfield and Nychka (1992) present frequency domain approaches for analyzing trends. Lettenmaier (1976) compares the ability of various nonparametric and parametric tests for detecting step and linear trends.

From an intuitive point of view, it may be instructive to consider an overall trend to consist of deterministic and stochastic components such that

$$\begin{array}{c}\textit{overall}\\ \textit{trend}\end{array} = \begin{array}{c}\textit{deterministic}\\ \textit{trend}\\ \textit{component}\end{array} + \begin{array}{c}\textit{stochastic}\\ \textit{trend}\\ \textit{component}\end{array}$$

In fact, this kind of interpretation forms the basis of the general intervention model in [19.5.8]. The *stochastic trend* is accounted for by the noise component which contains AR and MA operators to reflect the nonseasonal and seasonal correlation structures of the output and a white noise term for modelling what is left over when all the linear relationships contained in the output have been removed. Additionally, the noise component may require nonseasonal and seasonal differencing operators for modelling the nonstationary characteristics of a stochastic trend. Notice for the intervention model in [19.5.8] that when covariate series are available, they can also be incorporated into the model as extra information which usually increases the accuracy of all the parameter estimates in the overall intervention model and also removes effects upon the output which are not due to one or more interventions. Because it models the effects of one or more known interventions upon the output, the intervention term in [19.5.8] is dependent upon the time of occurrence of an intervention and therefore can be thought of as being a *deterministic component*. However, it should be kept in mind that this component is specifically designed to statistically describe the effects of known physical causes upon the output. Further discussions regarding deterministic and stochastic trends are presented in Sections 23.4.4 and 4.6.

When utilizing a statistical test, such as a specific kind of nonparametric test, to ascertain if there are trends in the data, one should always keep in mind exactly what the test is designed to detect and what are the *underlying assumptions* for the test. For example, due to the theoretical construction of a certain nonparametric test, it may only be designed for discovering the presence of an overall trend and may be incapable of distinguishing between stochastic and deterministic trend components. Further, as is the case for all of the nonparametric tests described in Section 23.3 for finding trends, nonparametric tests can only detect if a trend exists between the beginning and end of a time series and they cannot ascertain when the trends started due to external interventions. As a matter of fact, when only a small amount of data is available, the detection of the presence of trends is often all that one can realistically hope to achieve. Upon the collection of additional data, a more sophisticated procedure, such as intervention analysis, can be employed to describe more precisely the trend effects of known interventions.

When employing a nonparametric or parametric test to check for the presence of trends in a time series, there are different general approaches by which a given statistical test can be designed and executed. Therefore, a brief review of the *statistical testing* procedures consisting of hypothesis testing and significance testing is presented in the next section. Following a general discussion of available nonparametric tests in Section 23.3.1, specific nonparametric tests, which are used in the water resources literature for discovering trends in water quality and quantity time series (Hirsch et al., 1982; Hirsch and Slack, 1984; Van Belle and Hughes, 1984; Hirsch and Gilroy, 1985), are described in detail in Section 23.3.2. Of particular importance is the seasonal Mann-Kendall test (Hirsch et al., 1982) which can be used to test for the presence of a trend in each season of the year for a given data set. When testing for the presence of trends in, say, monthly data, one may wish to know whether each month should be checked separately or perhaps different groups of months should be tested together. Consequently, within Section 23.3.3 procedures are discussed for deciding upon how data should be grouped, and, in particular, the technique suggested by Van Belle and Hughes (1984) is discussed. For combining tests of hypotheses across seasons or groups of seasons, Fisher's (1970) method described in Section 23.3.4 is recommended. When dealing with water quality time series, often the effects of water quantity upon the water quality variables must be properly accounted for and in Section 23.3.5, regression analysis approaches for accomplishing this are described. In Section 24.3.2, the robust locally weighted regression smooth of Section 24.2.2 is utilized to allow for the effects of flow upon water quality when carrying out trend analysis studies of water quality time series measured in rivers. The Spearman partial rank correlation test is presented in Section 23.3.6 as a flexible nonparametric test for discovering trends in, say, a water quality variable measured over time when partialling out the effects of seasonality or riverflows upon the water quality variable. A nonparametric test is described in Section 23.3.7 for checking for the presence of a step trend caused by a known intervention in series measured at multiple stations. This test was devised by Hirsch and Gilroy (1985) and Crawford et al. (1983) and is related to the Mann-Whitney rank-sum test. In the final part of Section 23.3, procedures are presented for handling multiple censored data that are to be subjected to nonparametric trend testing. Within Section 23.4, the ACF (autocorrelation function) at lag one is suggested as a parametric test for finding trends. Using simulation experiments, the power of Kendall's tau (or equivalently the Mann-Kendall statistic) for detecting trends is compared to the power of this parametric statistic. The ACF at lag one is found to be more powerful than Kendall's tau for discovering purely stochastic trends while Kendall's tau is more powerful for finding purely deterministic trends. To demonstrate clearly the efficacy of utilizing various nonparametric tests and also some parametric methods in environmental impact assessment, practical applications are given in Section 23.5. In particular, nonparametric tests are utilized for discovering trends in water quality variables in Lake Erie caused by industrial development at the town of Nanticoke situated on the north shore of Lake Erie in the Canadian province of Ontario.

The nonparametric tests described in Chapter 23 are listed in Table 23.1.1 along with brief descriptions of their main purposes, equation numbers for the test statistics and the reference sources. Notice that the first six nonparametric tests are designed for checking for the presence of trends for a variety of situations and all of the trend tests are explained in Section 23.3. In addition to handling tied data, it is pointed out how the trend tests can be employed with censored time series. For the case of the seasonal Mann-Kendall test, procedures for taking care of correlation are also given. Finally, the last three nonparametric tests given in Table 23.1.1 are described in the three appendices and can be utilized for various useful tasks within a systematic data analysis study.

Table 23.1.1. Nonparametric tests described in Chapter 23.

NAMES	PURPOSES	TEST STATISTIC EQUATIONS	SOURCES
Nonseasonal Mann-Kendall	Determine if a time series contains a monotonic trend over time.	[23.3.1] and [23.3.4]	Mann (1945)
Seasonal Mann-Kendall	Find out if a seasonal time series contains an overall trend component.	[23.3.7], [23.3.8] and [23.3.11]	Hirsch et al. (1982), Hirsch and Slack (1984), Van Belle and Hughes (1984) and Lettenmaier (1988)
Aligned rank	Ascertain if a deseasonalized time series possesses a trend.	[23.3.23]	Sen (1968), Farrel (1980) and Van Belle and Hughes (1984)
Spearman's rho	Check if there is significant correlation between 2 variables X and Y. Can also be used as a trend test if one of the variables is time and the other is a sequence of observations.	[23.3.33] and [23.3.34]	Spearman (1904)
Spearman partial rank correlation	Determine the correlation between variables X and Y after the effects of Z upon X and Y are partialled out. Can be used to check if there is a trend in a series over time after seasonality is partialled out.	[23.3.35] and [23.3.36]	Based upon Spearman (1904)
Step trend	Find out if a known intervention causes significant step trends in series measured at multiple stations. Test is based on Mann-Whitney rank-sum test on grouped data.	[23.3.38] and [23.3.44]	Hirsch and Gilroy (1985) and Hirsch (1988)
Kendall rank correlation	Ascertain if two series X and Y are independent of one another. The Mann-Kendall trend test is a special case when one series is time and the other is sequential observations.	[A23.1.1]	Kendall (1975)
Wilcoxon signed rank	Check if two samples X and Y have the same median.	[A23.2.2] and [A23.2.3]	Wilcoxon (1945)
Kruskal-Wallis	Determine whether or not the distributions across k samples are the same. Can also be used to check if a time series possesses seasonality.	[A23.3.2] and [A23.3.3]	Kruskal and Wallis (1952)

As summarized in Table 1.6.4, three general approaches for carrying out trend analysis studies are presented in Sections 22.4, 23.5 and 24.3. Within each of these methodologies, appropriate exploratory and confirmatory data analysis tools can be employed, including the nonparametric techniques of Table 23.1.1. In fact, for the overall trend assessment procedures explained using environmental applications in Section 23.5 of this chapter and also Section 24.3, nonparametric trend tests have a key role to play. Besides the general approaches given in this book, other methodologies for trend assessment have been devised by researchers. For example, Montgomery and Reckhow (1984) suggest an overall systematic procedure for determining the presence or absence of trends in environmental data. Depending upon the characteristics of the data being analyzed, they suggest various nonparametric and parametric tests which can be used. Berryman et al. (1988) and Harcum et al. (1992) present systematic procedures for deciding upon which nonparametric tests to employ for detecting trends in water quality time series. Hipel and McLeod (1989) explain how both parametric and nonparametric models can be employed in trend assessment within the overall framework of exploratory and confirmatory data analyses. Hirsch et al. (1991) put forward procedures for selecting statistical methods to detect and estimate trends in water quality time series.

23.2 STATISTICAL TESTS

23.2.1 Introduction

Statistical testing can be carried out using nonparametric or parametric tests. However, a given statistical test, either nonparametric or parametric, can be designed for the purpose of hypothesis testing or significance testing. Cox and Hinkley (1974) present detailed descriptions of various kinds of hypothesis and significance tests which are briefly described in this section. The theory of tests of hypotheses was originally developed by Neyman and Pearson (1928, 1933) while significance testing is due largely to Fisher (1973).

23.2.2 Hypothesis Tests

Suppose one would like to determine whether or not a data set possesses a certain property. For example, one may wish to ascertain the existence or nonexistence of a certain kind of trend in a water quality time series. Typically, the *null hypothesis*, H_0, which is sometimes called the hypothesis under test, is that the population from which the sample data set is drawn, does not possess a specified property like a trend. The *alternative hypothesis*, H_1, which specifies a direction of departure from H_0, is that the data set does exhibit the property. In order to choose between H_0 and H_1 a *test statistic*, T, which is a function of the data set $\mathbf{X} = (x_1,x_2, \ldots ,x_n)$, is defined. When the hypothesis H_0 is true, the distribution of T must be known, at least approximately, so that the hypothesis test can be executed. By knowing the probability distribution of T, the probability of the sample statistic falling within or outside a given interval can be determined. As shown in Figure 23.2.1, for one sided and two sided tests, let t_r and t_l stand for the right and left values of T, respectively, for the two possible one sided tests, and let t'_r and t'_l define the right and left ends, respectively, of an interval in a two sided test. A chosen *significance level*, α, is the probability that the sample falls outside a specified range of values, given that H_0 is true. In practice, α is selected to have a value of 0.10, 0.05, or 0.01, although any appropriate value can be selected. For the one tailed or *one sided tests*, α represents the area in one of the tails of the distribution. In Figure 23.2.1a, $Pr(t \geq t_r) = \alpha$, and, consequently, if one were checking for the absence of an increasing trend, the hypothesis H_0 would be rejected if $t_x \geq t_r$ and hence H_1 would be accepted, where t_x is the sample or estimated value of T calculated using the sample X. Alternatively, H_0 would be accepted if $t_x < t_r$. The one tailed test in Figure 23.2.1b works in a similar manner. Suppose that H_0 indicates the absence of a decreasing trend. If $t_x \leq t_l$, then H_0 would be rejected and H_1 thereby accepted, whereas, when $t_x > t_l$, H_0 would be considered to be correct. A *two sided test* would be used in trend detection when one wishes to test for the presence of a trend which could be increasing or decreasing. For the case of the two sided test in Figure 23.2.1c, H_0 is accepted when $t'_l < t_x < t'_r$ and is rejected when $t_x \geq t'_r$ or $t_x \leq t'_l$.

Notice from Figure 23.2.1, that the probability of selecting H_0 when H_0 is true is $1-\alpha$, which is referred to as the *confidence level*. When executing a hypothesis test, two types of errors can arise. The probability of rejecting H_0 when H_0 is true is called a *type 1 error* or error of the first kind. From Figure 23.2.1, the probability of committing a type 1 error is α for both one sided and two sided tests. If, as a result of the same test statistic and a chosen significance level α, the hypothesis H_0 is accepted when it should be rejected, this is called a *type 2 error* or error of the second kind. Letting β represent the probability of committing a type 2 error, the probability of not making a type 2 error is $1-\beta$. The probability of rejecting H_0 when H_1 is true, or equivalently, the probability of not making a type 2 error, is called the *power* of the hypothesis test. If, for example, one were testing for the presence of a trend, the power, given by $1-\beta$, can be interpreted as the probability of detecting a trend when a trend is actually present in the data. In Section 23.4, the powers of two tests for detecting trends are compared using simulation studies for a number of different data generating models. When performing a hypothesis test, one of the four situations given in Table 23.2.1 can arise. Notice that two of the four outcomes to a hypothesis test result in either a type 1 or type 2 error. In Section 23.5, a variety of practical applications are presented using nonparametric statistics for hypothesis testing when checking for the presence of trends in water quality data.

In general, the power, $1-\beta$, and the confidence level, $1-\alpha$, are inversely related. Consequently, increasing the confidence level decreases the power and vice versa. For a specified significance level, α, the power of a test may be made greater by increasing the sample size.

23.2.3 Significance Tests

The type of significance test described here is what Cox and Hinkley (1974, Ch. 3) refer to as a *pure significance test*. As is also the case for hypothesis testing, when designing a significance test the null hypothesis, H_0, must be precisely formulated in terms of a probability distribution for the test statistic T. On the other hand, the major difference between the two tests is that a possible departure from H_0 in the form of an alternative hypothesis, H_1, is not rigorously defined for a significance test whereas it is assumed to be exactly known for a hypothesis test. Nonetheless, for a significance test it is necessary to have some general idea about the type of departure from H_0.

When performing a significance test, the acceptance or rejection of H_0 is decided upon in the same way as it is for a hypothesis test. Hence, as shown in Figure 23.2.1, one can perform either a one sided or two sided significance test. However, when H_0 is rejected only the general kind of departure from H_0 is known because there is no precise statement about an alternative hypothesis.

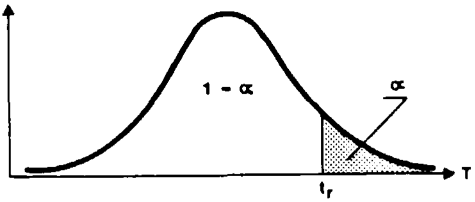


Figure 23.2.1a. One sided hypothesis test on the right.

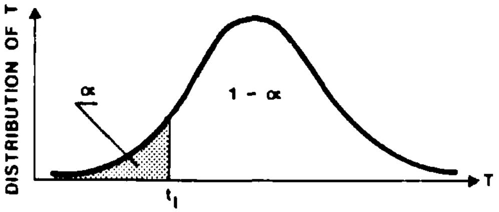


Figure 23.2.1b. One sided hypothesis test on the left.

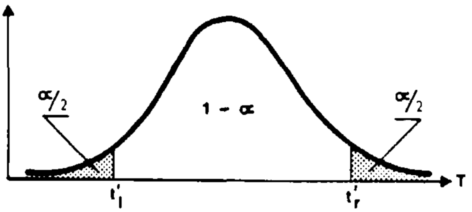


Figure 23.2.1c. Two sided hypothesis test.

Figure 23.2.1. Hypothesis tests.

Most of the diagnostic tests given in Chapter 7 constitute significance tests. For example, when one is checking whether or not the residuals of a stochastic model fitted to a time series are white, the null hypothesis may be that the residuals are white. Based upon an appropriate statistical test such as the Pormanteau statistic in [7.3.6], one can decide if H_0 should be accepted or rejected. When H_0 is rejected because the residuals are not white, the precise type of departure from H_0 is not explicitly defined. The residuals may be correlated for instance, because an ARMA(1,1) model should be fitted to the data instead of a an AR(1) model.

Table 23.2.1 Possible results of an hypothesis test.

	ACTION	
TRUE SITUATION	Accept H_0	Reject H_0
H_0 is true	No error (Pr = 1 - α) Confidence Level	Type 1 error (Pr = α)
H_1 is true	Type 2 error (Pr = β)	No error (Pr = 1 - β) Power

To avoid using too much statistical jargon and to enhance understanding by practitioners, hypothesis and significance tests are often not stated in a very formal manner when applying them to real data. As a matter of fact, statistical tests can be of assistance in the design of informal statistical tools for use in exploratory data analysis (see problem 22.5 in the previous chapter).

23.3 NONPARAMETRIC TESTS

23.3.1 Introduction

As noted in Section 23.1, a nonparametric test is also commonly referred to as a *distribution free or distribution independent method.* This is because no assumptions are made about the specific kind of distribution that the samples follow. The only restriction is that the samples come from the same basic population. Furthermore, a nonparametric test tends to be quite simple in design and easy to understand. In terms of time series analysis, most nonparametric tests can be used with both evenly and unevenly spaced observations.

Different types of data are available for use with statistical tests. The kinds of measurements are usually referred to as *measurement scales* or *systems of measurement*. From weakest to strongest, the four types of measurement scales recognized by Stevens (1946) are the nominal, rank (also called ordinal), interval, and ratio scales. In the *nominal scale* of measurement, numbers or other appropriate symbols are used for classifying objects, properties or elements into categories or sets. For example, when classifying objects according to colour, any number can be selected as a name for a given colour.

In the *rank or ordinal scale*, objects are ranked or ordered on the basis of the relative size of their measurements. A wine taster, for instance, may rank his wines from most to least desirable where a wine having a larger number assigned to it is more preferred than one with a smaller number. However, the amount by which one wine is preferred or not preferred over another is not specified. When analyzing a real world dispute using the conflict analysis approach referred to in Section 1.5.3, only ordinal preference information is assumed and hence each decision maker must rank the possible states or scenarios in the conflict from most to least preferred.

Besides the relative ordering of measurements used in the ordinal scale, the *interval scale* of measurement takes into account the size of the interval between measurements. An informative example of an interval scale is the common scales by which temperatures are usually measured. When using either the Fahrenheit or Celsius scale, a zero point and a unit distance (i.e., one degree of temperature) must be specified. Besides the Fahrenheit or Celsius scales, one could easily define any other temperature scale by stipulating the zero point and the degree unit. In other words, the principle of interval measurement is not violated by a change in scale or location or both.

To convert x degrees Celsius to y degrees Fahrenheit one uses the equation

$$y = \frac{9}{5}x + 32$$

When employing the interval scale, ratios have no meaning and one cannot, for instance, state that 10°C (50°F) is twice as warm as 5°C (41°F). Although the ratio of the two temperatures is $\frac{10}{5} = 2$ when using the Celsius scale, in the Fahrenheit scale the ratio is $\frac{50}{41} = 1.22$. Because there is a true or natural zero point in the Celsius scale, the concept of a ratio makes sense in this scale. When transforming from one *ratio scale* to another, it is only necessary to multiply one of the scales by a constant. Thus, for example, to transform x kilometres to y miles one uses the equation

$$y = 0.62x$$

One can say, for instance, that 10 km (6.2 miles) is twice as far as 5 km (3.1 miles) because $\frac{10}{5} = \frac{6.2}{3.1} = 2$. Other examples of ratio scales include temperature in degrees Kelvin, weight measured in kg or pounds, and time expressed in hours, minutes and seconds.

Most nonparametric methods are designed for use with data expressed in a nominal or ordinal scale. Because each scale of measurement possesses all of the properties of a weaker measurement scale, statistical methods requiring a weaker scale can be used with stronger scales. Consequently, time series observations which are always expressed using either an interval or a ratio scale, can be subjected to nonparametric testing. Most parametric methods can only be used with values given in an interval or ratio scale and cannot be utilized to analyze data belonging to a nominal or ordinal scale. Due to the foregoing and other reasons, Conover (1980, p. 92) defines a statistical method as being *nonparametric* if it satisfies at least one of the following criteria:

1. The method can be used with data possessing a nominal scale of measurement.
2. The method can be employed with data having an ordinal scale of measurement.
3. The method may be used with data having an interval or ratio scale of measurement where the probability distribution function of the random variable generating the data is either unspecified or specified except for an infinite number of unknown parameters.

In an *environmental impact assessment study*, usually the investigators have a fairly clear idea, at least in a general sense, of what they want to accomplish. For example, they may wish to ascertain if increased industrialization has significantly lowered the water quality of a large lake in the industrialized region. If data are not already available, a major task would be to design a suitable data collection scheme (see Sections 1.2.3 and 19.7). Assuming that observations are available for a range of water quality variables at different locations in the lake, a challenging problem is to select the most appropriate set of statistical methods that can be used in an optimal fashion for detecting and modelling trends in the data. Besides uncovering and modelling trends in water quality variables at a single site, statistical methods could be used to model the relationships among variables and trends across sites in the lake. In a large scale environmental impact assessment study, it is often necessary to use a variety of both nonparametric and parametric methods (see the applications in Sections 23.5, 22.3, 22.4 and 24.3).

As discussed in the introduction to Part X and also Sections 22.1, 22.3, as well as 1.2.4, when executing a data analysis study it is recommended to carry out *exploratory data analysis* followed by confirmatory data analysis. Usually simple graphical methods are employed at the exploratory data analysis stage for visually detecting characteristics in the data such as trends and missing values (see Section 22.3). Both nonparametric and parametric tests can be employed for hypothesis and significance testing during *confirmatory data analyses*. Because of the preponderance and proliferation of statistical methods, it is not surprising that a great number of statistical textbooks have been published and a significant number of papers have been printed in journals regarding the development and application of statistical methods (see Section 1.6.3). In fact, at least two major encyclopediae on statistics are now available (Kruskal and Tanner, 1978; Kotz et al., 1988) and a number of informative handbooks (see, for instance, Sachs (1984)) and dictionaries (Kendall and Buckland, 1971) on statistics have been written.

To assist in choosing the best statistical methods to use in a given study, the techniques can be classified according to different criteria. In an introductory paper to an edited monograph on time series analysis in water resources, Hipel (1985), for example, classifies time series models according to specified criteria. For the case of nonparametric tests, Conover (1980) presents a useful chart at the start of his book for *categorizing nonparametric tests* according to the kind of sample, hypothesis being tested, and type of measurement involved (nominal, ordinal and interval). Keep in mind that a test designed for a weaker type of measurement can also be used with stronger measurements. Consequently, all of the tests listed under nominal measurements can be used with both ordinal and interval data. Furthermore, the tests given below ordinal measurements can also be used to analyze interval measurements.

In the remainder of Section 23.3, nonparametric tests which are especially useful for detecting trends in water quality time series are described in detail. More specifically, the first six nonparametric trend tests listed in Table 23.1.1 are defined in Section 23.3 and other useful nonparametric procedures are also discussed. Fortunately, these tests can be modified for handling data sets having tied values as well as censored observations. Other topics in Section 23.3 include grouping seasons in a meaningful way in a trend detection study, procedures for combining trend tests across groups of seasons and adjusting water quality for riverflows.

The reader should keep in mind that in a trend assessment study, it is often necessary to employ a wide range of statistical methods. Although this text describes many useful parametric and nonparametric methods that are frequently used by environmental and water resources engineers, sometimes it may be necessary to refer to other texts and papers for a description of other methods. Besides the book of Conover (1980), other texts on nonparametric testing include contributions by Siegel (1956), Fraser (1957), Bradley (1968), Gibbons (1971, 1976), Hollander and Wolfe (1973), Puri and Sen (1971), Kendall (1975), and Lehman (1975). Gilbert (1987) describes a range of both nonparametric and parametric methods for use in environmental pollution monitoring. Because of the great importance of nonparametric methods in water resources and environmental engineering, the American Water Resources Association published a special monograph on this topic (Hipel, 1988). In the water quality applications in Section 23.5, ways in which a variety of nonparametric and parametric tests can be used for trend detection in a complex water quality study are explained. Other general methodologies for trend assessment are listed in Table 1.6.1 and also referenced at the end of Section 23.1.

23.3.2 Nonparametric Tests for Trend Detection

Introduction

In their paper, Van Belle and Hughes (1984) categorize nonparametric tests for detecting trends into two main classes. The one class is referred to as *intrablock methods* which are procedures that compute a statistic such as Kendall's tau for each block or season and then sum these to produce a single overall statistic (Hirsch et al., 1982; Hirsch and Slack, 1984). The second set of nonparametric tests are called *aligned rank methods*. These techniques remove the block effect from each datum, sum the data over the blocks and then create a statistic from these blocks (Van Belle and Hughes, 1984). The foregoing two classes of techniques are designed for detecting monotonic trends or changes (gradual or sudden) during some specified time interval but unlike the parametric technique of intervention analysis in Chapter 19 and Section 22.4 they are not intended for exploring the hypothesis that a certain type of change has occurred at some prespecified time due to a known external intervention.

Intrablock Methods

Because the tests of Hirsch et al. (1982) and Hirsch and Slack (1984) are based upon earlier work of Mann (1945) and Kendall (1975), the initial research is described first.

Mann-Kendall Test: Mann (1945) presented a nonparametric test for randomness against time which constitutes a particular application of Kendall's test for correlation (Kendall, 1975) commonly known as the *Mann-Kendall or the Kendall t test*. Letting $x_1,x_2,\ldots,x_n$, be a sequence of measurements over time, Mann (1945) proposed to test the null hypothesis, H_0, that the data come from a population where the random variables are independent and identically distributed. The alternative hypothesis, H_1, is that the data follow a monotonic trend over time. Under H_0, the *Mann-Kendall test statistic* is

$$S=\sum_{k=1}^{n-1}\sum_{j=k+1}^{n} sgn(x_j-x_k) \qquad [23.3.1]$$

where

$$sgn(x)=\begin{cases}+1, & x>0\\ 0, & x=0\\ -1, & x<0\end{cases}$$

Kendall (1975) showed that S is asymptotically normally distributed and gave the mean and variance of S, for the situation where there may be ties in the x values, as

$$E[S]=0$$

$$Var[S]=\left\{n(n-1)(2n+5)-\sum_{j=1}^{p} t_j(t_j-1)(2t_j+5)\right\}/18 \qquad [23.3.2]$$

where p is the number of tied groups in the data set and t_j is the number of data points in the jth tied group.

When using [23.3.1], a positive value of S indicates that there is an upward trend in which the observations increase with time. On the other hand, a negative value of S means that there is a downward trend. Because it is known that S is asymptotically normally distributed and has a mean of zero and variance given by [23.3.2], one can check whether or not an upward or downward trend is significantly different from zero. If the S is significantly different from zero, based upon the available information H_0 can be rejected at a chosen significance level and the presence of a monotonic trend, H_1, can be accepted. For a general review of how to execute a hypothesis test, the reader can refer to Section 23.2.2.

The exact distribution of S for $n \leq 10$ was derived by both Mann (1945) and Kendall (1975). They showed that even for small values of n, the normality approximation is good provided one employs the standard normal variate Z given by

$$Z=\begin{cases} \dfrac{S-1}{[Var(S)]^{1/2}} & \text{, if } S>0 \\ 0 & \text{, if } S=0 \\ \dfrac{S+1}{[Var(S)]^{1/2}} & \text{, if } S<0 \end{cases} \qquad [23.3.3]$$

The statistic S in [23.3.1] is a count of the number of times x_j exceeds x_k, for $j>k$, more than x_k exceeds x_j. The maximum possible value of S occurs when $x_1<x_2< \cdots <x_n$. Let this number be called D. A statistic which is closely related to S in [23.3.1] is *Kendall's tau* defined by

$$\tau=\frac{S}{D} \qquad [23.3.4]$$

where

$$D=\left[\frac{1}{2}n(n-1)-\frac{1}{2}\sum_{j=1}^{p} t_j(t_j-1)\right]^{1/2}\left[\frac{1}{2}n(n-1)\right]^{1/2}$$

When there are no ties in the data, [23.3.4] collapses to

$$\tau=\frac{S}{\frac{1}{2}n(n-1)}=\frac{S}{\binom{n}{2}} \qquad [23.3.5]$$

Due to the relationship between τ and S in [23.3.4], the distribution of τ can be easily obtained from the distribution of S. If there are no ties in the data, the algorithm of Best and Gipps (1974) can be employed to obtain the exact upper tail probabilities of Kendall's tau, or equivalently S, for $n \geq 2$.

The Mann-Kendall trend test is, in fact, a special case of the *Kendall rank correlation test.* Kendall (1975) developed the Kendall rank correlation test for ascertaining if two series are independent of one another. This test is described in Appendix A23.1.

Seasonal Mann-Kendall Test: Hirsch et al. (1982) defined a multivariate extension of the Mann-Kendall statistic in [23.3.1] for use with seasonal data, and, as noted by Van Belle and Hughes (1984), their test possesses some similarities to tests proposed by Jonckheere (1954) and Page (1963). Although the statistic of Hirsch et al. (1982), is valid for use with data where there may be missing values and also ties, assume for the present that the time series X consists of a complete record sampled over n years where there are m seasons per year such that X is given by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

The null hypothesis, H_0, is that for each of the m seasons the n observations are independent and identically distributed while the alternative hypothesis is there is a monotonic trend. Let the matrix of ranks be denoted by

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ R_{21} & R_{22} & \cdots & R_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ R_{n1} & R_{n2} & \cdots & R_{nm} \end{bmatrix}$$

where the n observations for each season or column in R are ranked among themselves. Hence, the rank of x_{jg}, which is the jth data point in the gth season, is

$$R_{jg} = [n + 1 + \sum_{i=1}^{n} sgn(x_{jg} - x_{ig})]/2 \qquad [23.3.6]$$

and each column of R is a permutation of $(1,2,\ldots,n)$. The *Mann-Kendall test statistic for the gth season* is (Hirsch et al., 1982)

$$S_g = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_{jg} - x_{ig}), \quad g = 1,2,\ldots,m \qquad [23.3.7]$$

Similar to the situation of S in [23.3.1], S_g is asymptotically normally distributed where

$$E[S_g] = 0$$

$$Var[S_g] = \sigma_g^2 = n(n-1)(2n+5)/18 \qquad [23.3.8]$$

Kendall's tau statistic for the gth season is defined as

$$\tau_g = \frac{S_g}{\frac{1}{2}n(n-1)} = \frac{S_g}{\binom{n}{2}} \qquad [23.3.9]$$

Because τ_g is simply a multiple of S_g, the distribution of τ_g can be obtained from the distribution of S_g. In particular, τ_g is asymptotically normally distributed where

$$E[\tau_g] = 0$$

$$Var[\tau_g] = \frac{1}{\left[\frac{1}{2}n(n-1)\right]^2}\left[\frac{n(n-1)(2n+5)}{18}\right] \qquad [23.3.10]$$

Since it is arithmetically more convenient to deal with S_g rather than τ_g and also Hirsch et al. (1982) use mainly S_g rather than τ_g in their research, the statistic S_g is utilized in the rest of this section.

Following Hirsch et al. (1982), the *seasonal Mann-Kendall test statistic* is

$$S' = \sum_{g=1}^{m} S_g \qquad [23.3.11]$$

which is asymptotically normally distributed where

$$E[S'] = 0$$

$$Var[S'] = \sum_{g=1}^{m}\sigma_g^2 + \sum_{\substack{g,h \\ g\neq h}}\sigma_{gh} \qquad [23.3.12]$$

Using [23.3.8], $\sigma_g^2 = Var[S_g]$ can be calculated as well as $\sigma_{gh} = cov(S_g S_h)$. For the situation where each season is independent of each of the other seasons, the second summation in [23.3.12] is zero and

$$Var[S'] = \sum_{g=1}^{m}\sigma_g^2 \qquad [23.3.13]$$

As is done in [23.3.3] for the Mann-Kendall test, for $n \leq 10$ the standard normal deviate Z' should be calculated as

$$Z' = \begin{cases} \dfrac{S'-1}{[Var(S')]^{1/2}} & \text{, if } S' > 0 \\ 0 & \text{, if } S' = 0 \\ \dfrac{S'+1}{[Var(S')]^{1/2}} & \text{, if } S' < 0 \end{cases} \qquad [23.3.14]$$

When utilizing [23.3.14], Hirsch et al. (1982) demonstrate that the normal approximation is quite accurate even for data as small as $n = 2$ and $m = 12$.

To handle *missing values*, $sgn(x_{jg} - x_{ig})$ is defined to be zero if either x_{jg} or x_{ig} is missing. Letting n_g be the number of nonmissing observations for season g, equation [23.3.6] is modified as

$$R_{jg} = [n_g + 1 + \sum_{i=1}^{n} sgn(x_{jg} - x_{ig})]/2 \qquad [23.3.15]$$

Consequently, the ranks of the known observations remain unchanged and each missing observation is assigned the average or midrank $(n_g + 1)/2$. As is the case when there are no missing observations, equation [23.3.7] is used to calculate S_g and following [23.3.8] the variance of S_g is determined using

$$\sigma_g^2 = n_g(n_g - 1)(2n_g + 5)/18 \qquad [23.3.16]$$

S' and its variance are determined using [23.3.11] and [23.3.12], respectively, where each σ_{gh} is zero.

For a *censored time series*, in which some data are reported to be less than a *detection limit*, arbitrarily fix the affected data at some constant value which is less than the limit of detection. Because nonparametric tests are based upon ranks instead of magnitudes, all censored values are interpreted as sharing the same rank which is less than the rank of all uncensored observations. Additionally, this means that handling censored data is equivalent to dealing with *ties*. Assuming, for the moment, that there are no missing values, the ranked data containing ties can be calculated using [23.3.6], which automatically assigns to each of t tied values the average of the next t ranks. Following this, S_g can be determined utilizing [23.3.7] where, similar to the situation in [23.3.2], the variance is

$$Var[S_g] = \sigma_g^2 = [n(n-1)(2n+5) - \sum_{j=1}^{p} t_j(t_j - 1)(2t_j + 5)]/18 \qquad [23.3.17]$$

where n is the number of years of data, p is the number of tied groups for the data x_{ig}, $i = 1,2,\ldots,n$, in season g, and t_j is the size of the jth tied group. The seasonal Mann-Kendall statistic is calculated using [23.3.7] while its variance is determined by utilizing [23.3.12] where all the σ_{gh} are zero. When there are both tied data (due to "tied" censored data and ties of actual observations) and missing values, the modifications described in this and the previous paragraph must be combined. Finally, a general description of censored data is presented in Section 23.3.8.

Another problem which can arise when using any of the nonparametric tests in this section, is how to summarize information when there are *several values for a specified season* in a given year. Van Belle and Hughes (1984, p. 135) suggest four possible approaches for accomplishing this. One method is to adjust the seasonal length so that there is not more than one observation per season. This, of course, would mean having a smaller seasonal length, such as biweekly instead of monthly. A second approach is to take a single value which is closest to the center of the season in a given year and to ignore the other information. In many applications, a third, and more reasonable method to follow, is to simply replace the set of values by the median or mean before calculating the test statistic. As a matter of fact, many water quantity and quality records which are collected on a daily basis are often released in reports as average monthly values. Fourthly, an alternative to calculating a mean or median within a given season and year, is to consider these values as tied in the time index and then to compute the test statistic along with a modified variance. Van Belle and Hughes (1984, p. 135) present the formulae for carrying this out.

Recall that for the intervention analysis applications of seasonal data in Chapter 19 and Section 22.4, it is suggested that an intervention may affect each season or groups of seasons in different manners. For example, when modelling the impacts of reservoir operation upon the average monthly flows of the South Saskatchewan River in Section 19.2.5, it is suspected that the seasonal mean levels would increase during certain months and decrease at other times. To accurately model an upward or downward step trend as well as the change in magnitude of the mean for each season, a separate intervention term is incorporated into the intervention model for each month. In a similar manner, one should examine carefully how S_g in [23.3.7] behaves for each season. Only if the same type of trend, such as an upward trend, is detected in each season, will the overall seasonal Mann-Kendall test statistic in [23.3.11] have any meaning. In other words, S' should only be calculated for a group of seasons which are expected to behave in a certain manner where hypothesis testing is done separately for this group. A more detailed discussion of this problem is presented in Section 23.3.3 where approaches are presented for *combining tests of hypotheses across seasons.*

Correlated Seasonal Mann-Kendall Test: In practice, environmental data are usually correlated and not independently distributed as is assumed in the previous section. For instance, when dealing with average monthly phosphorous levels, the phosphorous observations in one month may be significantly correlated with values in the preceding one or more months. This means that in order to employ the seasonal Mann-Kendall tests in [23.3.7] and [23.3.11], one must be able to estimate all the σ_{gh} in [23.3.12].

Based upon research by Dietz and Killeen (1981), Hirsch and Slack (1984) explain how σ_{gh} can be estimated. Assuming, for the moment, that there are no ties or missing values, a consistent estimation for σ_{gh} is

$$\hat{\sigma}_{gh} = K_{gh}/3 + (n^3 - n)r_{gh}/9 \qquad [23.3.18]$$

where

$$K_{gh} = \sum_{i=1}^{n-1}\sum_{j=1}^{n} sgn[(x_{jg} - x_{ig})(x_{jh} - x_{ih})] \qquad [23.3.19]$$

$$r_{gh} = \frac{3}{n^3 - n}\sum_{i,j,k} sgn[(x_{jg} - x_{ig})(x_{jg} - x_{kh})] \qquad [23.3.20]$$

For the situation where there are no ties and no missing values, the statistic r_{gh} is *Spearman's* *correlation coefficient* for seasons g and h (Conover, 1980; Lehman, 1975). If there are no missing observations, equation [23.3.18] can be written as

$$\hat{\sigma}_{gh} = [K_{gh} + 4\sum_{j=1}^{n} R_{jg}R_{jh} - n(n+1)^2]/3 \qquad [23.3.21]$$

where each R_{jg} is determined using [23.3.6].

To employ the correlated seasonal Mann-Kendall test when there are no ties or missing data, S_g, $g = 1,2, \cdots ,m$, and S' are determined using [23.3.7] and [23.3.11], respectively. Following the estimation of the variance of each S_g and σ_{gh} using [23.3.8] and [23.3.18], respectively, the variance of S' can be calculated using [23.3.12]. As can be done for the seasonal Mann-Kendall test, a separate hypothesis can be formulated for each season or each group of seasons, and, when appropriate, based upon S' an overall hypothesis can be made. For a given season or group of seasons, seasonal data are thought to be independently distributed since it is assumed that a particular data point is not correlated with data occurring one or more years before during the same season. Consequently, the null hypothesis, H_0, for a given season is the data are independent and identically distributed while H_1 is the existence of a monotonic trend in that season. For the overall correlated seasonal Mann-Kendall statistic, S', the null hypothesis is that the data are correlated and identically distributed while H_1 is the presence of a monotonic trend. Because the mean, variance and distribution of each S_g and also S' are known, hypothesis testing can be executed. As is the case for the seasonal Mann-Kendall test, when employing the correlated seasonal Mann-Kendall test, the standard normal deviate in [23.3.14] should be calculated for $n \leq 10$.

To use the correlated seasonal Mann-Kendall test with *missing values*, the procedure described for the seasonal Mann-Kendall test is used. The ranks of the data for season g are determined using [23.3.15], and then S_g and its variance are calculated using [23.3.7] and [23.3.8], respectively. The correlated seasonal Mann-Kendall statistic is determined using [23.3.11]. To estimate the variance for S' using [23.3.12], equation [23.3.18] or equivalently [23.3.21] must be appropriately modified in order to estimate σ_{gh} for substitution into [23.3.12]. The K_{gh} term is determined as before by using [23.3.19]. However, in the presence of missing values, r_{gh} takes a new form which causes the revised version of [23.3.21] to be

$$\hat{\sigma}_{gh} = [K_{gh} + 4\sum_{j=1}^{n} R_{jg}R_{jh} - n(n_g+1)(n_h+1)]/3 \qquad [23.3.22]$$

where n_g is the number of observations for season g and n_h stands for the number of measured values for season h.

As in the previous section, *censored data* are handled as ties where data reported as being less than a limit of detection are assigned a constant value which is less than the limit of detection. Assuming that there are no missing values, the ranks containing ties can be calculated using [23.3.6], which automatically assigns to each of t tied values the average of the next t ranks. Next, S_g can be determined by utilizing [23.3.7], while the variance of S_g can be calculated using [23.3.17]. The correlated seasonal Mann-Kendall statistic, S', is determined using [23.3.11]. To employ [23.3.18] or [23.3.21] for estimating σ_{gh}, midranks are used in assigning the values of R_{jg}. Hence, if there were t_j censored values, they would all have a rank of $t_j(t_j - 1)/2$. Following this, the variance of the correlated seasonal Mann-Kendall statistic can be determined using [23.3.12]. When there are both tied data, which may be due to censored data and ties of actual observations, and missing values, the alterations outlined in this and the previous paragraph must be combined. Lettenmaier (1988) proposes a technique called the covariance eigenvalue method to handle correlation among seasons. He uses simulation experiments to compare the power of his seasonal trend test to those of Hirsch and Slack (1984) as well as Dietz and Killeen (1981). Utilizing simulation experiments, Loftis et al. (1991a,b) find that Lettenmaier's (1988) technique works well with serially correlated data. Zetterqvist (1988) also proposed a seasonal Mann-Kendall trend test to take care of autocorrelation among data in different seasons, where the observations within each season are assumed to be independent.

El-Shaarawi and Niculescu (1992) extend the Kendall Tau test for handling correlation in nonseasonal data when the underlying process is MA(1) or MA(2). Moreover, they develop a test for use with seasonal data having non-zero correlations between successive seasons and years.

Seasonal Kendall Slope Estimator: The intrablock statistics discussed thus far, are designed for detecting the presence but not the location or magnitude of a trend in a time series. To estimate the magnitude of a trend, Hirsch et al. (1982) suggest an extension to the seasonal case of the method proposed by Theil (1950) and Sen (1968). In particular, the seasonal Kendall slope estimator of Hirsch et al. (1982) expresses the magnitude of a trend as a slope which means change of the series per unit time. When sufficient data are available, the technique of intervention analysis discussed in Chapter 19 and Section 22.4, constitutes a much more powerful procedure for accurately estimating the magnitude of trends caused by one or more known interventions. As shown in Figures 19.2.3 and 19.2.4 and explained in Section 19.2.2, the intervention component contained within the overall intervention model for modelling a trend can be designed so that the geometric shape of the trend is correctly modelled. Other trend detection techniques include the exploratory data analysis graphs of Section 22.3, the change-detection statistics referred to in Sections 19.2.3 and 24.2.1, and the robust locally weighted regression smooth of Section 24.2.2.

The seasonal Kendall slope estimator is defined to be the median of the differences, expressed as slopes, of the ordered pairs of data points that are compared in the seasonal Mann-Kendall test. The computational algorithm for defining the seasonal Kendall slope estimator, Sl, is as follows. Calculate

$$d_{ijk} = (x_{ij} - x_{ik})/(j - k)$$

for all (x_{ij}, x_{ik}) pairs $i = 1, 2, \cdots, m$, where $1 \leq k < j \leq n_i$ and n_i is the number of known values in the ith season. The slope estimator Sl is the median of the d_{ijk} values. As noted by Hirsch et al. (1982), the slope estimator Sl is closely related to S' in [23.3.11]. If $S' > 0$ then $Sl \geq 0$ ($Sl > 0$ if one or no $d_{ijk} = 0$) and if $S' < 0$, then $Sl \leq 0$ ($Sl < 0$ if one or no $d_{ijk} = 0$). The reason for these relationships between S' and Sl is that S' is equivalent to the number of positive d_{ijk} values minus the number of negative d_{ijk} values and Sl is the median of the d_{ijk} values.

As pointed out by Hirsch et al. (1982), because the median of the d_{ijk} values is used to define SI, the estimator is quite resistant to the presence of extreme observations in the data. Further, since the slopes are always computed between values that are integer multiples of the seasonal length m, the slope estimator is unaffected by seasonality. If, for example, the magnitude of the slope were thought to be different for each season, the slope could be estimated separately for each season. In situations where groups of seasons are suspected of having the same magnitude for the slope estimator, the slope estimator could be separately applied to each group of seasons. A computer program for calculating the seasonal Mann-Kendall statistic and the seasonal Kendall slope is listed by Smith et al. (1982) and also Crawford et al. (1983).

Aligned Rank Methods

In addition to the intrablock methods, the aligned rank techniques constitute nonparametric approaches for checking for the presence of trends in a data set. However, unlike the intrablock methods which can be used with incomplete records, the aligned rank techniques are designed for use with evenly spaced observations for which there are no missing values. One particular kind of aligned rank method is described in this section.

Suppose that a time series, **X**, consists of a complete record sampled over n years where there are m seasons per year. Hence, following the notation suggested by Van Belle and Hughes (1984), the data can be displayed in the following fashion:

		Season						
		1	2	.	.	.	m	mean
	1	x_{11}	x_{12}	.	.	.	x_{1m}	$x_{1.}$
	2	x_{21}	x_{22}	.	.	.	x_{2m}	$x_{2.}$
	.	.	.				.	.
Year	.	.	.				.	.
	.	.	.				.	.
	n	x_{n1}	x_{n2}	.	.	.	x_{nm}	$x_{n.}$
	mean	$x_{.1}$	$x_{.2}$	.	.	.	$x_{.m}$	$x_{..}$

When a dot is used to replace a subscript, this indicates the mean taken over that subscript. Therefore,

$$x_{.j} = \sum_{i=1}^{n} x_{ij}/n$$

is the mean of the jth season and

$$x_{i.} = \sum_{j=1}^{m} x_{ij}/m$$

is the mean for the ith year.

Based upon the work of Sen (1968), Farrell (1980) proposed the following procedure to test for the presence of trends which is also described by Van Belle and Hughes (1984).

1. The data are deseasonalized by subtracting the seasonal average from each data point as in [13.2.2] in Section 13.2.2. However, because each season is suspected of containing a trend, one may question the validity of deseasonalization which assumes a constant mean for each season.
2. The nm deseasonalized data points are ranked from 1 to nm. When t values are tied, simply assign the average of the next t ranks to each of the t tied values. The matrix of ranks for the deseasonalized data can be written as:

		Season						
		1	2	.	.	.	m	mean
	1	R_{11}	R_{12}	.	.	.	R_{1m}	$R_{1.}$
	2	R_{21}	R_{22}	.	.	.	R_{2m}	$R_{2.}$
	.	.	.				.	.
Year	.	.	.				.	.
	.	.	.				.	.
	n	R_{n1}	R_{n2}	.	.	.	R_{nm}	$R_{n.}$
	mean	$R_{.1}$	$R_{.2}$	.	.	.	$R_{.m}$	$R_{..}$

3. The average ranks for each season and year are obtained. As is the case for the given set of data, **X**, the mean over a subscript is indicated by a dot. Accordingly,

$$R_{.j} = \sum_{i=1}^{n} R_{ij}/n$$

is the average rank for the jth season while

$$R_{i.} = \sum_{j=1}^{m} R_{ij}/m$$

is the average rank for the ith year.

4. For use in hypothesis testing, the following statistic is calculated.

$$T = \left[\frac{12m^2}{n(n+1)\sum_{j=1}^{m}\sum_{i=1}^{n}(R_{ij} - R_{.j})^2}\right]^{1/2} \cdot \left[\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)\left(R_{i.} - \frac{nm+1}{2}\right)\right] \qquad [23.3.23]$$

In reality, T is the slope of the regression of $R_{i.}$ against i for $i = 1,2,\ldots,n$, standardized by the square root of the residual error given by

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(R_{ij}-R_{.j})^2/[m(n-1)]$$

The null hypothesis, H_0, is that the data are identically independently distributed and, hence, possess no trends. The alternative hypothesis, H_1, is that the data have trends. For large samples, T approaches normality with a mean of zero under H_0 and a variance of unity. Consequently, the test statistic T is approximately a standardized normal variable and, as explained in Section 23.2.2, either a one sided or two sided statistical test can be executed.

If a data set is incomplete, the *missing observations* must be estimated before the aligned rank method can be employed. Depending upon how many data points are missing, a suitable data filling procedure from Sections 22.2, 19.3 or 18.5.2, can be selected for obtaining an estimated evenly spaced time series. Farrell (1980) suggests estimating missing data using a least squares approach. Subsequent to the data filling, the aligned rank test statistic in [23.3.23] can be calculated in order to carry out an hypothesis test.

An alternative approach to estimating missing data, is to adjust the seasonal length until there is at least one observation per season for each year (Van Belle and Hughes, 1984). For example, suppose that one were examining weekly data for which there are quite a few weeks containing no data points. However, when a monthly series is considered for the same data set, there is at least one observation per month of every year. By calculating the mean value for each month within a given year, a time series of evenly spaced monthly values can be created. The aligned rank method can then be applied to this monthly series.

As noted earlier, *ties* of the deseasonalized values can easily be handled using the aligned rank method. If there are *censored values* where some data are reported to be below a limit of detection, the affected data can be arbitrarily fixed at some constant value which is less than the limit of detection. A drawback of this approach is that the estimates of the yearly and seasonal means will be biased. Therefore, it should be used only if the relative number of censored values is not too great. Nevertheless, subsequent to determining the ranks of the deseasonalized data, the statistic in [23.3.23] can be calculated.

Comparison of Intrablock and Aligned Rank Methods

As pointed out by Van Belle and Hughes (1984), both the aligned rank method and the seasonal Mann-Kendall test are based upon the same model given by

$$x_{ij}=\mu+a_i+b_j+e_{ij},\quad i=1,2,\ldots,n\text{ , and } j=1,2,\ldots,m \qquad [23.3.24]$$

where μ is the overall mean, $\mathbf{a}=\{a_1,a_2,\ldots,a_n\}$, is the yearly component, $\mathbf{b}=\{b_1,b_2,\ldots,b_m\}$, is the seasonal component and

$$\sum_{i=1}^{n}a_i=\sum_{j=1}^{m}b_j=0.$$

The e_{ij} is the noise term which is independently distributed. For both nonparametric tests, the null hypothesis is that the yearly component is zero and hence

$$H_0\text{: } \mathbf{a}=\mathbf{o}$$

The alternative hypothesis is

$$H_1\colon\ a_1 \le a_2 \le a_3 \ \cdots\ \le a_n \ \text{and/or}$$

$$a_1 \ge a_2 \ge a_3 \ \cdots\ \ge a_n$$

with at least one strict inequality. Accordingly, both approaches are testing for a monotonic trend over the years where the trend is not necessarily linear. A trend within each year could be considered as part of the seasonal trend **b**.

Using the results of Puri and Sen (1971), Van Belle and Hughes (1984) show that the aligned rank test is always more powerful than the seasonal Mann-Kendall test and that the difference is greater for smaller numbers of years of data. As mentioned before, when there are missing data, the intrablock tests can be used without having to estimate the missing observations. However, in order to use the aligned rank method with an unevenly spaced time series, the missing observations must be estimated prior to applying the technique. Finally, using simulation experiments, Taylor and Loftis (1989) find that the correlated seasonal Mann-Kendall test is more powerful than its competitors, including an aligned rank method, for detecting trends.

23.3.3 Grouping Seasons for Trend Detection

As was pointed out in the discussion included with the seasonal Mann-Kendall test, an intervention may affect each season or groups of seasons in different ways. Based upon a physical understanding of the problem and using exploratory data analysis procedures such as the time series plots described in Section 22.3.2 and the box-and-whisker graphs of Section 22.3.3, one can decide how seasons should be grouped together. For instance, as explained in Section 19.5.4, a physical comprehension of the problem and exploratory data analyses make one suspect that a forest fire caused the spring flows of the Pipers Hole River in Newfoundland, Canada, to increase immediately after the fire and to gradually attenuate over the years back to their former levels as the forest recovered. However, during other seasons of the year the fire did not cause any trends in the time series after the fire. By employing the technique of intervention analysis, this behaviour is rigorously confirmed and accurately modelled in Section 19.5.4.

Nonparametric testing can be executed in a fashion similar to the general approach used in intervention analysis studies. In order to *classify seasons into groups* where seasons within each group possess the same kind of trend, one procedure is to rely upon a *physical understanding of the problem* and the output from *exploratory data analyses*. The *Kruskal-Wallis test* (Kruskal and Wallis, 1952) can also be used to test for the presence of seasonality and decide upon which seasons are similar (see Appendix A23.3). The statistic defined for the nonparametric test being used to detect trends can be calculated separately for each group of seasons to ascertain if a certain kind of trend is present within the group. Output from the nonparametric tests may suggest other ways in which the seasons should be grouped and then the statistics can be calculated for the new grouping of the seasons.

Consider, for example, how the seasonal Mann-Kendall test can be used when seasons are grouped according to common patterns recognized in trends. The Mann-Kendall statistic, S_g, for each season can be calculated using [23.3.7] and the variance, σ_g^2, of S_g can be determined using [23.3.8]. Suppose that one of the groups of seasons consists of seasons in the set represented by G. Then the *seasonal Mann-Kendall statistic* for the seasons in group G is calculated as

$$S_G = \sum_{g \in G} S_g \qquad [23.3.25]$$

The variance of S_G is then determined as

$$Var[S_G] = \sum_{g \in G} \sigma_g^2 \qquad [23.3.26]$$

where the expected value of each S_g and also S_G is zero. Because S_G is asymptotically normally distributed, hypothesis testing can be done to see if the seasons in group G possess a common trend. Recall that a significantly large positive value of S_G would indicate an increasing trend, while a significantly large negative value of S_G would mean there is a decreasing trend. Further, for small samples one should employ [23.3.14] to calculate the standard normal deviate where S_G replaces S' in [23.3.14]. In a similar fashion, groups of seasons could be considered when employing the correlated seasonal Mann-Kendall statistic and also the aligned rank method.

For deciding upon how common or homogeneous trends should be grouped, Van Belle and Hughes (1984) suggest employing a *homogeneity test* (Fleiss, 1981, Ch. 10) which is commonly used in the study of cross-classified data. This test is closely related to the seasonal Mann-Kendall test of Section 23.3.2. Van Belle and Hughes (1984) propose that the grouping or homogeneity test should be used as a preliminary test for checking for the homogeneity of trends and, thereby, classifying the seasons into groups where each group possesses a common trend. Subsequent to this, an intrablock statistic such as the seasonal Mann-Kendall statistic or the aligned rank statistic can be calculated for each group to ascertain if there is a significantly large common trend in the group.

For use in the homogeneity test of Van Belle and Hughes (1984), the statistic Z_g^2 is defined as

$$Z_g^2 = S_g^2/Var[S_g] \qquad [23.3.27]$$

where the Mann-Kendall statistic, S_g, for the gth season is given in [23.3.7] and its variance, $Var[S_g]$ is presented in [23.3.8]. Because Z_g is asymptotically normally distributed, Z_g^2 approximately follows a chi-squared distribution with one degree of freedom. As in [23.3.9], Kendall's tau for the gth season is related to S_g by the expression

$$\tau_g = 2S_g / \left[n_g(n_g - 1) \right] \qquad [23.3.28]$$

where n_g is the number of data points in the gth season. The null hypothesis, H_0, is there is no trend in the gth season and it can be written as H_0: $\tau_g = 0$ or, equivalently, H_0: $S_g = 0$.

Notice that because the square of S_g is used in [23.3.27], the sign of S_g is eliminated in the calculation of Z_g^2. Because a positive or negative sign for S_g indicates an increasing or decreasing trend, respectively, one should make sure that different kinds of trends are not being combined when Z_g^2 is summed across seasons. Suppose that a physical appreciation of the problem in conjunction with output from preliminary data analyses indicate that seasons in a set labelled G should be included within one group. This group, for example, may stand for the group of summer seasons where there is an increasing trend in each summer season and, therefore, an increasing trend for the entire group G. For the group G, the overall *homogeneity test statistic* is

$$\chi_G^2 = \sum_{g \in G} Z_g^2 \qquad [23.3.29]$$

which is approximately χ^2 distributed with $|G|$ degrees of freedom where $|G|$ is the number of seasons in group G. The null hypothesis, H_0, is that there is no trend across the seasons in G and, therefore, the Kendall τ or the Mann-Kendall statistic for each season in G is zero. If, at a selected level of significance, the statistic calculated using [23.3.29] is significantly different from zero, one can reject H_0 and conclude that there is an overall trend across the seasons in the group G. The statistic in [23.3.29] is, of course, separately calculated for each grouping of seasons where every season is a member of one of the groups.

The foregoing approach is one way of deciding upon how seasons should be grouped. Other approaches are given by Van Belle and Hughes (1984), Fleiss (1981) and Zar (1974).

In the general situation, one may wish to examine trends in various water quality variables across an entire river basin or other appropriate geographical entities for which there is a set of locations where data are collected. Consequently, for a given variable one must not only ascertain how seasons should be grouped at a single station but also how data can be grouped across stations. For each water quality variable, one procedure is to employ a physical understanding of the problem and exploratory data analyses executed for data collected at each site to decide upon which seasons and site locations should be included in G used for calculating a group statistic such as the Mann-Kendall statistic in [23.3.25] and the homogeneity statistic in [23.3.29]. Based upon the χ^2 statistic in [23.3.29], Van Belle and Hughes (1984) propose a method for *grouping data for a given variable across seasons and sites*.

As noted earlier, one can use a statistical test such as the nonparametric *Kruskal-Wallis test* (see Appendix A23.3) to determine whether or not a given time series contains seasonality. If the data are not seasonal, then, of course, all of the observations fall under one group. One can then employ the nonseasonal Mann-Kendall test in [23.3.1] or [23.3.5] to check for trends. However, one should not employ the seasonal Mann-Kendall trend test when seasonality is not present. This would certainly result in a loss of power in the trend test.

23.3.4 Combining Tests of Hypotheses

One may calculate a test statistic such as the Mann-Kendall statistic S_g in [23.3.7] for each season of the year when examining a seasonal time series. Alternatively, by following one of the procedures described in Section 23.3.3, one may join seasons together and calculate a separate statistic, such as S_G in [23.3.25], for each group of seasons. Whatever the case, following the determination of the test statistic and associated significance level for each season or each group of seasons, one may wish to then combine tests of hypotheses across seasons or groups of seasons. For explanation purposes, suppose that one wants to calculate a separate Mann-Kendall statistic for each of the seasons and then combine tests of hypotheses across seasons in order to arrive at an overall hypothesis test. As noted by Littell and Folks (1971), several authors have considered the problem of combining independent tests of hypotheses. Using the exact Bahadur relative efficiency (Bahadur, 1967), Littell and Folks (1971) compare four methods of combining independent tests of hypotheses. The methods they compare are Fisher's (1970) method, the mean of the normal transforms of the significance levels, the maximum significance level, and the minimum significance level. Although none of the tests is uniformly more powerful than the others, according to the Bahadur relative efficiency, Fisher's method is the most efficient of the four.

Let the observed significance level of a test of hypothesis be denoted by SL_i. For example, because the distribution of S_g in [23.3.7] is known for a given data set, one can calculate SL_g for S_g where $g = 1,2, \ldots, m$. Because of the relationship between Kendall's τ_g and S_g in [23.3.9], SL_g would be the same for both τ_g and S_g in season g. When there are m independent tests, Fisher (1970, p. 99) shows that

$$-2\sum_{i=1}^{m} lnSL_i \approx \chi^2_{2m} \qquad [23.3.30]$$

For the situation where SL_g is the observed significance level for S_g or, equivalently, τ_g in the gth season, the null hypothesis would be that the data for all of the seasons considered in the test come from a population where the random variables are independent and identically distributed. The alternative hypothesis is that the data across the seasons follow a monotonic trend over time. If, for example, the magnitude of the observed chi-squared variable calculated using [23.3.30] were larger than the tabulated χ^2_{2m} value at a chosen significance level, one would reject the null hypothesis. In [10.6.7] within Section 10.6.4, Fisher's combination method is used to demonstrate that ARMA models fitted to geophysical time series statistically preserve the Hurst coefficient and hence provide an explanation for the Hurst phenomenon.

23.3.5 Flow Adjustment of Water Quality Data

In [22.4.5] of Section 22.4.2, an intervention model is presented for describing the effects of cutting down a forest upon a seasonal water quality time series. Notice in [22.4.5] that the response variable, which represents the water quality variable under consideration, is dependent upon a number of different components written on the right hand side of [22.4.5]. Of particular interest is the fact that the riverflows are included as a covariate series in the intervention model and the manner in which the flows stochastically affect the output is modelled by the specific design of the transfer function for the riverflows. Accordingly, the influence of water quantity upon a given water quality variable is realistically and rigorously accounted for by including the flows as an input series to a water quality intervention model.

When employing a nonparametric test for checking for the presence of trends, a more accurate study can be executed if the impacts of water quantity upon water quality are properly accounted for. For a long time, scientists have known that many water quality variables are correlated with river discharge (Hirsch et al., 1982; Langbein and Dawdy, 1964; Johnson et al., 1969; Smith et al., 1982). Consider, for example, the case of total phosphorous which can have a rather complex dependence upon riverflows (Reckhow, 1978; Hobbie and Likens, 1973; Borman et al., 1974). As mentioned by Smith et al. (1982), at base flow conditions in certain watersheds, much of the phosphorous may be due to point-source loadings and, hence, a decrease in flow would cause an increase in phosphorous concentrations. Alternatively, in some river basins the occurrence of a massive rainstorm over a basin may cause the erosion and transport of organic and inorganic materials which carry large amounts of phosphorous and, therefore, the resulting increases in riverflows may be combined with increased phosphorous levels. Consequently, for a given river it is important to have a physical understanding of the type of relationship which exists between a given water quality variable and runoff. In some river basins, more than one physical process may take place where each process is a function of the quantity of riverflow. As noted by Harned et al. (1981), streamflow is the single largest source of variability in water quality data.

Hirsch et al. (1982) and Smith et al. (1982) suggest a general procedure by which the different effects of water quantity upon a water quality variable can be modelled. The purpose of their procedures is to develop a time series of *flow adjusted concentrations (FAC)* for the water quality variable under consideration which can then be tested for trends using appropriate nonparametric tests described in Section 23.3 and elsewhere. Depending upon the physical characteristics of the problem being studied, an appropriate filter can be designed to obtain the FAC series. In general, an equation for determining concentrations may have the form

$$X = f(Q) + \varepsilon \qquad [23.3.31]$$

where Q is the flow, $f(Q)$ gives the functional relationship of the flows upon the water quality variable under consideration and also contains the model parameters, ε is the noise, and X represents the concentration. For the situation where increased flows causes dilution of the water quality variable, $f(Q)$ may have one of the following forms (Hirsch et al., 1982):

$$f(Q) = \lambda_1 + \frac{\lambda_2}{Q}$$

$$f(Q) = \lambda_1 + \frac{\lambda_2}{1 + \lambda_3 Q}$$

where λ_i is the ith parameter. If increased precipitation and hence runoff increase the concentration of a water quality variable, it may be reasonable to model $f(Q)$ as

$$f(Q) = \lambda_1 + \lambda_2 Q + \lambda_3 Q^2$$

As explained by Hirsch et al. (1982) and Smith et al. (1982), regression analysis can be employed to determine which form of $f(Q)$ is most appropriate to use. Given that a significant relationship can be found using regression analysis, the FAC for year i and season j is calculated as

$$w_{ij} = x_{ij} - \hat{x}_{ij}$$

where w_{ij} is the estimated FAC, x_{ij} is the observed concentration and $\hat{x}_{ij}$ is the estimated concentration which is determined using linear regression with the best form of $f(Q)$ in [23.3.31]. The FAC series can then be subjected to nonparametric tests in order to check for trends.

When obtaining the FAC series, one is in fact using a parametric procedure to properly filter the original observations for use with a nonparametric test. An advantage of this approach is that it can be used with unevenly spaced time series. A drawback is that in regression analysis the noise term is assumed to be white. If sufficient data are available so that an estimated evenly spaced series can be obtained, the technique of intervention analysis constitutes a single parametric approach which is a much more flexible and powerful procedure for modelling water quality series. Note only does the intervention model account for the stochastic effects of flows upon the water quality variable but it also rigorously models the forms and magnitudes of trends caused by known interventions. Indeed, if it is suspected that the manner in which flows affect the water quality variable depends upon the season of the year, appropriate dynamic components can be designed to model this behaviour. Further, as explained for the intervention model in [22.4.5] and elsewhere in Chapter 19, any number of input series and trends can be modelled and the noise can be described by a correlated process such as an ARMA model.

In Section 24.3.2 a general methodology is presented for analyzing trends in water quality series measured in rivers. To remove the effects of riverflow upon a given water quality variable, the robust locally weighted regression smooth described in Section 24.2.2 is employed as one of the steps in the overall procedure. Subsequently, the Spearman partial rank correlation trend test described in Section 23.3.6 and other appropriate trend tests are employed for formally testing for the presence of trends in the water quality series. Other approaches for compensating for discharge when evaluating trends in water quality one provided by Harned et al. (1981). Bodo and Unny (1983) explain how stratified sampling can improve the estimation of load-discharge relationships.

23.3.6 Partial Rank Correlation Tests

Introduction

The previous subsection deals with the problem of adjusting water quality data in a river for the impacts of flow before checking for the presence of a trend in the water quality data. To eliminate, hopefully, the effects of flow, various regression models can be used, as described in Section 23.3.5 and also Section 24.3.2 in the next chapter.

The removal of flow effects from a water quality time series when testing for a trend, is part of a more general statistical problem. More specifically, when studying the dependence between two variables X and Y, one may wish to know if the correlation between X and Y is caused by the correlation of both X and Y with a third variable Z. For instance, one may want to find out if a possible trend in a water quality variable, as manifested by the correlation of the water quality variable over time, is independent of riverflows. Hence, one would like to remove or *partial out* the influence of water quantity when testing for a trend in the water quality variable over time. Another example for which eliminating certain effects is desirable, is when one wishes to check for trend in a seasonal water quality variable against time when the seasonality has been partialled out.

The objective of this section is to present a nonparametric trend test in which undesirable effects can be removed. In particular, the Spearman partial rank correlation test (McLeod et al., 1991) is suggested as a useful trend test for employment in environmental engineering. Because this test utilizes some definitions used in the Spearman's rho test, this latter test is first described. Additionally, the Spearman partial rank correlation test is compared to the seasonal Mann-Kendall test of Section 23.3.2 as well as the Kendall partial rank correlation coefficient (Kendall, 1975). Applications of the Spearman partial rank correlation test to a seasonal water quality time series are given in Section 24.3.2.

Spearman's Rho Test

In 1904, Spearman introduced a nonparametric coefficient of rank correlation denoted as ρ_{XY} which is based upon the squared differences of ranks between two variables. Spearman's rho can be employed as a nonparametric test to check whether or not there is significant correlation between two variables X and Y.

Let the sample consist of a bivariate sample (x_i,y_i) for $i = 1,2,\ldots,n$, where n is the sample size. Suppose that the values of the X variable are ranked from smallest to largest such that the rank of the smallest value is one and that of the largest value is n. Let $R_i^{(X)}$ represent the rank of the X variable measured at time i. Likewise, the values of the Y variable can be ranked and $R_i^{(Y)}$ can represent the value of the rank for the Y variable at time i. The sum of the squared differences of the ranks is

$$S(d^2) = D^2 = \sum_{i=1}^{n}(R_i^{(X)} - R_i^{(Y)})^2 \qquad [23.3.32]$$

Spearman's rho is then defined for the case where there are no ties in X and Y as

$$\rho_{XY} = 1 - \frac{6S(d^2)}{n^3 - n} \qquad [23.3.33]$$

When the two rankings for X and Y are identical, then $\rho_{XY} = 1$ whereas $\rho_{XY} = -1$ when the rankings of X and Y are in reverse order.

If some values of X or Y are tied, these values are simply assigned the average of the ranks to which they would have been assigned. Let p be the number of tied groups in the X data set where t_j is the number of data points in the jth tied group. Likewise, let q be the number of tied groups in the Y sequence where u_j is the number of observations in the jth tied group. Then the formula for calculating ρ_{xy} when there are ties in either or both time series is (Kendall, 1975, p. 38, Equation 3.8)

$$\rho_{XY} = \frac{\frac{1}{6}(n^3 - n) - S(d^2) - \frac{1}{12}\sum_{j=1}^{p}(t_j^3 - t_j) - \frac{1}{12}\sum_{j=1}^{q}(u_j^3 - u_j)}{\left[\left\{\frac{1}{6}(n^3 - n) - \frac{1}{6}\sum_{j=1}^{p}(t_j^3 - t_j)\right\}\left\{\frac{1}{6}(n^3 - n) - \frac{1}{6}\sum_{j=1}^{q}(u_j^3 - u_j)\right\}\right]^{1/2}} \qquad [23.3.34]$$

When using ρ_{XY} in a statistical test to check for the absence or presence of correlation, the null hypothesis, H_0, is that there is no correlation, For large samples, ρ_{XY} is distributed as $N\left(0, \frac{1}{n} - 1\right)$ where n is the sample size. The alternative hypothesis, H_1, is that there is correlation between the X and Y variables.

By letting one of the variables represent time, Spearman's rho test can be interpreted as a *trend test*. In particular, replace (x_i,y_i) by (t,x_t) for which $t = 1,2,\ldots,n$, and x_t consists of $x_1,x_2,\ldots,x_n$. Equations [23.3.32] to [23.3.34] can then be employed to calculate a statistic for use in a trend test. If, for example, the estimated value of ρ_{XY} is significantly different from zero, then one can argue that time and the X variable are significantly correlated, which in turn means there is a trend.

Spearman Partial Rank Correlation Test

When examining dependence between two variables X and Y, the question arises as to whether or not the correlation between X and Y is due to the correlation of each variable with a third variable Z. For example, one may wish to ascertain if an apparent trend in a water quality variable, as reflected by the correlation of the water quality variable over time, is independent of seasonality. Therefore, one would like to eliminate or *partial out* the effects of seasonality when testing for a trend in the water quality variable over time.

The purpose of the Spearman partial rank correlation test presented in this section is to determine the correlation between variables X and Y after the effects of Z upon X and Y separately are taken into account and are, therefore, removed. The notation used to represent this type of partial correlation is $corr(XY/Z)$ or $\rho_{XY.Z}$.

Let the sample consist of a trivariate sample (x_i, y_i, z_i) for $i = 1,2,\ldots,n$, where n is the sample size. As is also done for the Spearman's rho test, suppose that the values of the X variable are ranked from smallest to largest such that the rank of the smallest value is one and that of the largest value is n. Let $R_i^{(X)}$ represent the rank of the X variable at time i. Likewise, the values of the Y and Z variables can be ranked separately to produce $R_i^{(Y)}$ and $R_i^{(Z)}$, respectively.

The test statistic for the Spearman partial rank correlation test is calculated using

$$\rho_{XY.Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{(1-\rho_{XZ}^2)^{1/2}(1-\rho_{YZ}^2)^{1/2}} \qquad [23.3.35]$$

Each rho term on the right hand side of the above equation is calculated using [23.3.33] when there are no ties in any variable or [23.3.34] where there are ties.

Under the null hypothesis, there is no correlation between X and Y when the effects of Z are partialled out. To test the null hypothesis that $E(\rho_{XY.Z}) = 0$, one calculates

$$t = \frac{(n-2)^{1/2}\rho_{XY.Z}}{(1-\rho_{XY.Z}^2)^{1/2}} \qquad [23.3.36]$$

where $\rho_{XY.Z}$ is defined in [23.3.35]. Under H_0, t follows a student t distribution on $(n-2)$ degrees of freedom (Pitman's approximation), which is the same as ρ_{XY} in [23.3.33] and [23.3.34]. The alternative hypothesis, H_1, is there is correlation between X and Y, after accounting for the influence of Z separately upon X and Y. If, for example, the SL for the test statistic were calculated to be very small and less than 0.05, one could reject the null hypothesis that X and Y are not correlated when Z is partialled out. When the X, Y and Z variables represent a given water quality time series, time and seasonality, respectively, then a significantly large value of the Spearman partial rank correlation coefficient means that there is a trend in the series over time when seasonality is removed. Besides checking for trend after removing seasonality, the Spearman partial rank correlation test can be used for other purposes. For example, it can be extended to take into account correlation when testing for the presence of a trend.

The partial Spearman correlation test can be used with data for which there are missing values, ties and one level of censoring, either on the left or right. If the data are multiply censored, one can use the expected rank vector approach of Hughes and Millard (1988) before applying the test, which is discussed in Section 23.3.8. Some theoretical developments and simulation experiments for the partial Spearman correlation test are reported by Valz (1990).

Comparison to the Seasonal Mann-Kendall Test

Besides the Spearman partial rank correlation test of this section, recall that the seasonal Mann-Kendall test of Section 23.3.2 can also be used to check for trend in a seasonal time series. However, the Spearman partial rank correlation test has advantages over the seasonal Mann-Kendall test including:

1. simulation experiments demonstrate that it has more power,
2. it is more flexible and can be extended, for example, to take into account correlation,
3. and it provides an estimate of the magnitude of the trend through the coefficient in [23.3.35]. For example, when one is using the Spearman partial rank correlation coefficient, to check for a trend in a variable X over time when seasonality is partialled out (see explanation in last subsection), larger positive and negative values indicate bigger upward and downward trends over time, respectively.

Kendall Partial Rank Correlation Coefficient

Another coefficient for determining the correlation between two variables X and Y when the effects of Z upon each of these variables is taken into account is the Kendall partial rank correlation coefficient (Kendall, 1975, Ch. 8). However, because of the way the statistic is defined, it possesses some serious theoretical drawbacks. The end result is the distribution of the statistic is not known so that it cannot be used in hypothesis testing (Valz, 1990).

Before describing in more detail specific disadvantages, some notation is required. Let τ_{XY}, τ_{ZY} and τ_{XZ} represent the Kendall rank correlation coefficient between X and Y, Z and Y and X and Z, respectively. Also, let $\tau_{XY.Z}$ be the Kendall partial rank correlation statistic to determine the rank correlation between X and Y, taking into account the effects of Z upon these variables. Finally, let $\rho'_{XY.Z}$ be the Pearson partial rank correlation coefficient. For equations defining these statistics, the reader can refer to Kendall (1975).

The disadvantages of using the Kendall partial rank correlation coefficient include:

1. If the X, Y and Z variables are multivariate normally distributed and the Pearson partial rank correlation coefficient $\rho'_{XY.Z} = 0$, it can be shown using simulation and also from the relationship

$$E(\tau_{XY.Z}) = \frac{2}{\pi}\sin^{-1}\rho'_{XY.Z} \qquad [23.3.37]$$

that $\tau_{XY.Z}$ is different from zero. This does not occur with the Spearman partial rank correlation coefficient.

2. The variance of the estimate for $\tau_{XY.Z}$ depends on τ_{XY}, τ_{ZY} and τ_{XZ}. The analogous undesirable property does not hold for the Pearson or Spearman partial rank correlation coefficient.
3. On intuitive grounds, it appears that the Kendall partial rank correlation coefficient does not eliminate in a linear way the effects of Z.

Because of the foregoing disadvantages, the Kendall partial rank correlation coefficient is not used in applications by the authors.

23.3.7 Nonparametric Test for Step Trends

The statistical tests described in Sections 23.3.2 to 23.3.4 as well as 23.3.6 are designed solely for discovering trends in a data set. These techniques do not take into account when trends may have started due to external interventions. In this section, the nonparametric test of Hirsch and Gilroy (1985) and Crawford et al. (1983) is described for ascertaining if a known intervention causes a significantly large step trend for series measured at multiple stations. This test is closely related to the Mann-Whitney rank-sum test. As noted by Hirsch and Gilroy (1985), their test does not depend on parametric model assumptions, does not require complete data sets and is resistant to the effects of outliers. Nevertheless, if an evenly spaced data set can be estimated from incomplete records, the technique of intervention analysis could be used for accurately modelling trends. Recall from Chapter 19 and Section 22.4 that intervention analysis can be employed to estimate the exact magnitude of a step trend or, for that matter, many other types of trend created by known interventions.

In their paper, Hirsch and Gilroy (1985) examine the detectability of step trends in the monthly rate of atmospheric deposition of sulphate. A downward step trend in the sulphate levels would be due to a sulphate emission control program which came into effect at a known date. Prior to employing the statistical test, it is recommended that the original data set be appropriately filtered to remove unwanted sources of variation. When dealing with a water quality variable such as phosphorous, the technique described in Section 23.3.4 is a procedure for filtering the phosphorous data so the effects of water quantity upon water quality are taken into account. For the case of removing unwanted variation from a time series describing the rate of atmospheric deposition of sulphate, Hirsch and Gilroy (1985) present a specific type of filtering to remove the portion of the variance in sulphate loading rates which is due to the variance in precipitation rates and also to the variance in the seasonally varying mean values. In particular, the filtered sulphate loading series consists of the residuals obtained from the regression of logarithmic sulphate loadings on the logarithmic precipitation series.

Hirsch and Gilroy (1985) apply the nonparametric test for step trends to filtered sulphate loading series available at n_s measuring stations. The nonparametric test is the Mann-Whitney rank-sum test on grouped data which is described by Bradley (1968, p. 105). To apply the test to the same physical variable measured across seasons at n_s sites, the following steps are adhered to:

1. If it is not advisable to test the original series, obtain a filtered series by utilizing an appropriate filter. As just noted, for the case of sulphate loading series, one may wish to employ the filter presented by Hirsch and Gilroy (1985) while one may wish to use one of the filters described in Section 23.3.5 when dealing with certain kinds of water quality data.

2. Calculate the Mann-Whitney rank-sum statistic and other related statistics for data which are grouped according to season and station. In particular, by letting the subscripts i, j, and k represent the year, season and station, respectively, for group jk (season j and station k), the *Mann-Whitney rank-sum statistic* is

$$W_{jk} = \sum_{i=1}^{n_1} R_{ijk} \qquad [23.3.38]$$

where n_1 and n_2 are the number of years before and after the known intervention, respectively, for which data are collected, and R_{ijk} is the rank over the entire n years of the filtered data, X_{ijk}, $i = 1,2, \cdots ,n = n_1 + n_2$, which are ranked for the jth season and kth station. The statistic in [23.3.38] is determined for every season at each of the stations. If one assumes the null hypothesis that there is no trend in group jk, the statistic W_{jk} has the expectation of

$$\mu = n_1(n_1 + n_2 + 1)/2 \qquad [23.3.39]$$

and variance of

$$\sigma^2 = n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)/m \qquad [23.3.40]$$

where m is the number of seasons. Note that the previous three equations assume that all groups have the same record length. However, the test described in this section can be used with data sets where the numbers of data points are different across groups. Although extra notation could be used to allow for varying record lengths across the groups, for simplicity of explanation the foregoing and upcoming equations are explained for the situation where each group has the same record length.

If the data are independent, the mean and variance of the sum of the Mann-Whitney rank-sum statistics across all the groups can be easily determined. In particular, the mean and variance of $\sum_{j=1}^{m} \sum_{k=1}^{n_s} W_{jk}$ are given by $m \cdot n_s \cdot \mu$ and $m \cdot n_s \cdot \sigma^2$, respectively.

The variance in [23.3.40] is based upon the assumption that the data and, hence, the W_{jk} are independently distributed. Hirsch and Gilroy (1985) describe the following approach for estimating σ_{jk}^2 when the data are correlated. The covariance between W_{jk} and W_{gh} is given by

$$C(W_{jk},W_{gh}) = \sigma^2 \rho(X_{ijk},X_{igh}) \qquad [23.3.41]$$

where $\rho(X_{ijk},X_{igh})$ is the rank correlation between data in season j station k and data in season g station h. The variance of the sum of the W_{jk}'s is determined using

$$Var\left[\sum_{j=1}^{m} \sum_{k=1}^{n_s} W_{jk}\right] = \sum_{j=1}^{m} \sum_{g=1}^{m} \sum_{k=1}^{n_s} \sum_{h=1}^{n_s} C(W_{jk},W_{gh}) \qquad [23.3.42]$$

By assuming that the serial correlation of the ranks is lag one autoregressive and the same correlation coefficient can be used at all the stations, the estimation of the covariances can be greatly simplified. Based upon these assumptions, the estimated correlation coefficient, r_1, can be easily calculated. All of the ranks, R_{ijk}, except the last one, R_{nmk}, for each station $k = 1,2, \ldots , n_s$, are paired with the rank, $R_{i,j+1,k}$, of the succeeding filtered observations except when $j = m$ it is $R_{i+1,1,k}$. The product moment correlation coefficient of all of the pairs determines r_1. The covariances $C(W_{jk},W_{gk})$ between different seasons at the same station are estimated using

$$\hat{C}(W_{ik},W_{gk}) = \sigma^2 r_1 | g - i |$$

The covariances $C(W_{jk},W_{jh})$, $k \neq h$, for different stations and the same season are estimated as

$$\hat{C}(W_{jk},W_{jh}) = \sigma^2 r_0(k,h)$$

where $r_0(k,h)$ is the product moment correlation coefficient of the concurrent ranks

$$(R_{ijk},R_{ijh}),\ i = 1,2, \ldots, n;\ j = 1,2, \ldots, m$$

for stations k and h. For different stations and different seasons, the covariances $C(W_{jk},W_{gh})$, $j \neq g$, $k \neq h$, are estimated using

$$\hat{C}(W_{jk},W_{gh}) = \sigma^2 r_0(k,h) r_1 | g - i |$$

When considering the same season and station, the covariance $C(W_{jk},W_{jk})$ is simply the variance σ^2 given in [23.3.40]. Based upon the foregoing, to estimate the variance of the sum of the W_{jk} statistics in [23.3.42], the following expression is utilized

$$Var\left[\sum_{j=1}^{m}\sum_{k=1}^{n_s} W_{jk}\right] = \sum_{j=1}^{m}\sum_{g=1}^{m}\sum_{k=1}^{n_s}\sum_{h=1}^{n_s}\hat{C}(W_{jk},W_{gh}) \qquad [23.3.43]$$

3. Perform a hypothesis test to ascertain if there is a significantly large step trend in the time series due to a known intervention. The null hypothesis, H_0, is that there is no step trend while the alternative hypothesis is that there is a step trend. This test could be restricted to a certain group of seasons of the year if it were expected that the intervention only affected the seasons within that group. For example, a pollution spill may only influence certain physical variables when the temperature is above a certain level and, therefore, the data from the winter months may be excluded from the group. For the purpose of the test described here, it is assumed that a step trend may be formed for data in each season across all of the stations due to a single known intervention.

   The test statistic for checking the validity of the null hypothesis is

$$Z' = \left(\sum_{j=1}^{m}\sum_{k=1}^{n_s} W_{jk} - \mu_{jk}\right) / \sigma_{jk} \qquad [23.3.44]$$

where μ_{jk} and σ_{jk}^2 are the mean and variance, respectively, of

$$\sum_{j=1}^{m}\sum_{k=1}^{n_s} W_{jk}$$

For the situation where the filtered series and hence the W_{jk} are independent, the mean and variance are given by $m \cdot n_s \cdot \mu$ and $m \cdot n_s \cdot \sigma^2$, respectively, where μ and σ^2 are presented in [23.3.39] and [23.3.40], respectively. When the data are correlated, the mean is still given as $m \cdot n_s \cdot \mu$ but the variance is calculated using [23.3.43]. Because Z' is asymptotically normally distributed one can compare the estimated value of Z' to the value of a standard normal distribution at a selected significance level. If, for example, the estimated value of Z' were significantly different from zero, based upon the available evidence one could reject H_0 and thereby conclude that there is a significant step trend in the data. Because the W_{jk} in [23.3.38] are calculated for the ranks before the intervention, a negative value of Z' would indicate an upward step trend after the intervention whereas a positive value of Z' would indicate a downward step trend after the intervention. To demonstrate the efficacy of the aforesaid test for detecting a step trend caused by a known intervention, Hirsch and Gilroy (1985) perform simulation studies. For the situation where the data are independently distributed, Crawford et al. (1983) present a computer program to calculate the test statistic. Research on comparing statistical methods for estimating step trends and their use in sampling design is presented by Hirsch (1988).

23.3.8 Multiple Censored Data

Introduction

As noted in Section 23.3.2, often water quality data are reported as being less than a *detection level*. These observations are referred to as *censored data*. If a single limit of detection is used for a specified time series, the data is said to be *singly censored*. When there is more than one detection limit, the observations are *multiple censored*.

To apply a nonparametric test to singly censored data, the version of the test modified for use with ties can be employed. In Section 23.3.2, for instance, it is explained how the seasonal Mann-Kendall trend test can be applied to a time series with one detection limit by simply treating the censored observations as being tied. If, however, the detection limits vary within a time series and, therefore, the data are multiple censored, then one should follow other approaches in order to employ nonparametric tests.

Because multiple detection levels occur frequently in practice, the purpose of this section is to put this problem into perspective and point out procedures for handling multiple censored observations so that one can apply a given nonparametric test to the data set. In this way, practitioners will be able to make the most efficient use of the data available to them for estimating test statistics or parameters, even though the observations may possess the undesirable property of being multiple censored.

There are a variety of reasons as to why water quality and other kinds of data have multiple detection levels. As noted by authors such as Millard and Deverel (1988) and Helsel and Cohn (1988), these include:

1. The detection level changes because different methods are used to measure water quality samples at various time periods, either in the field or in the laboratory. For example, over time analytical methods may improve so that the detection levels are lowered.

2. To reduce costs, management may at different points in time request the use of cheaper measurement techniques which have higher detection levels.

3. A range of methods may be available for measuring a given water quality variable at any given time. However, each technique may have a range of the concentration of the variable for which it can provide the optimal measurement. Hence, each method has a different detection level.

4. Multiple detection levels can be caused by the process of dilution. For example, because of time constraints, a laboratory technician may adhere to a procedure whereby he can only have a specified maximum number of dilutions for any single sample. Since the detection limit is dependent upon the amount of dilution, this procedure may create multiple detection limits.

5. When data are sent to several agencies or laboratories for analyses, these organizations may have different reporting levels. Often environmental bodies such as the Environmental Protection Agency in the United States and Environment Canada are obligated to send their samples to many different private and government laboratories for analyses in order to treat everyone fairly. However, this may result in having multiple censored data.

As mentioned by Millard and Deverel (1988) as well as other authors in the field of water resources, an impressive array of techniques for handling censored data was originally developed within the areas of *survival analysis and life testing* (see, for example, Kalbfleisch and Prentice (1980)). Consequently, the basic censoring definitions and methods developed in these areas are outlined and then the censoring techniques that are suitable for use with water quality and other types of environmental data are pointed out in the next section. An attractive procedure to use with multiple censored data is the expected rank vector method first suggested for use in environmental engineering by Hughes and Millard (1988).

Censoring Definitions in Survival Analysis

Before defining censoring, first consider the meaning of *truncation*. A sample of data is said to be *truncated on the left* if only observations above a specified truncation point are reported. Likewise, a data set is *truncated on the right* when only measurements below a given truncation level are used. If, for example, a phosphorous sample is left truncated at 5 mg/*l*, then only the measurements that are greater than 5 mg/*l* would be reported.

A sample consisting of n observations is *singly censored on the left* if n_c of these measurements, where $n_c \geq 1$, are known only to fall below a *censoring level* c. The remaining $(n-n_c)$ uncensored observations would thus lie above the censoring or detection level and would be fully reported. A sample of n measurements is *multiple censored on the left* with m censoring levels if $n_{c1}, n_{c2}, \ldots,$ and n_{cm} observations are censored on the left at levels $c_1, c_2, \ldots,$ and c_m, respectively.

In a similar fashion, one can also define singly or multiple censored observations on the right. For instance, a sample of n observations is *singly censored on the right* if $n_{c'}$ of these observations are known only to fall above a specified censoring level c' while the remaining $(n-n_{c'})$ observations are reported exactly.

One can further characterize censoring according to type I and type II censoring. A singly censored sample of n measurements constitutes *type I censoring on the left* if a given censoring level c_1 is specified in advance and values below c_1 are only reported as less than c_1. Likewise, a singly censored sample of n observations arises from *type I censoring on the right* when a specified censoring level is fixed in advance and observations lying above c_1 are simply reported as being greater than c_1.

When there is *type II censoring on the left*, only the r largest observations of a sample of size n, where $1 \leq r < n$, are reported, and the remaining $(n - r)$ measurements are known to lie below the rth largest values.

For *type II censoring on the right*, only the r smallest measurements of a sample of size n, where $1 \leq r < n$, are reported, while the remaining $(n-r)$ observations are known to lie above the rth smallest value. As an example of type II censoring, consider a situation where one is determining the failure times of n electronic components which are started at the same time. This experiment is stopped under type II censoring after r of the components have failed.

When dealing with environmental data, only some of the definitions developed in survival analysis and life testing are required for practical purposes. In particular, *environmental time series having detection limits almost always fall under the category of type I left censoring for either single or multiple censoring*. Within the field of survival analysis, usually right censored data are encountered. Fortunately, many statistical techniques developed for use with right censored data can be converted for use with data censored on the left. The reader may wish to refer to texts by authors such as Kalbfleisch and Prentice (1980), Lee (1980) and Miller (1981) for a description of statistical censoring techniques used in survival analysis and life testing.

Multiple Censoring in Environmental Engineering

In the area of environmental research, work has been carried out for estimating parameters when the data sets are singly censored (Kushner, 1976; Owen and DeRouen, 1980; Gilbert and Kennison, 1981; Gilliom et al., 1984; Gleit, 1985; Gilliom and Helsel, 1986; Gilliom and Helsel, 1986; El Shaarawi, 1989; Porter and Ward, 1991). For the case of the seasonal Mann-Kendall trend test of Section 23.3.2, Gilliom et al. (1984) demonstrate the effects of censoring with one detection limit upon the power of the test.

Although less research has been carried out in the environmental area for handling multiple censored data, some valuable contributions have been made. Helsel and Cohn (1988) use Monte Carlo methods to compare eight procedures for estimating descriptive statistics when the data are multiple censored. They show that the adjusted maximum likelihood technique (Cohn, 1988) and the plotting position method (Hirsch and Stedinger, 1987) perform substantially better then what are called simple substitution methods. Millard and Deverel (1988) discuss nonparametric tests for comparing medians from two samples, explain how multiple censored data can be handled when using these tests, and then employ Monte Carlo studies to compare the tests.

An innovative approach to extend the nonseasonal and seasonal Mann-Kendall trend tests of Section 23.3.2 for use with multiple censored data is the method proposed by Hughes and Millard (1988) which is referred to as a *tau-like test for trend in the presence of multiple censoring points*. The first step is to assign an average rank for each observation and thereby obtain a rank vector, by taking into account all permissible combinations of ranks in the presence of multiple censoring. Second, after the expected ranks are obtained for each observation in order to get the overall *expected rank vector*, a standard linear rank test can be applied to the expected rank vector. Hughes and Millard (1988) show in detail how this approach is carried out with the Mann-Kendall trend test.

To explain how the procedure of Hughes and Millard (1988) works in practice, consider the situation presented below in Table 23.3.2 where measurements are available at four points in time. The symbol X^- indicates a left censored observation at detection level X. For this simple example, notice that there are the two detection levels: 10 and 4. Below the vector of observations are the possible three rank vectors that could occur for the data set. In each rank vector, the observations are ranked from 1 to 4, where the smallest value is assigned a 1 and the largest observation a 4. Notice that the observation at time 4 is always the largest value and, therefore, is assigned a rank of 4 in each of the three rank vectors. However, depending upon how far below a detection level the unknown actual observation may fall, one can obtain the possible rankings as shown in the table. The expected rank vector listed in the last row of Table 23.3.1 simply gives the average rank across the three possible rank vectors at each point in time.

Table 23.3.1. Hypothetical example for calculating the expected rank vector.

	Time t			
	1	2	3	4
Observation X_t	10^-	9	4^-	18
Possible	3	2	1	4
Rank	2	3	1	4
Vectors	1	3	2	4
Expected Rank Vector	2	2.7	1.3	4

To apply a nonparametric test to data having multiple censoring levels, one simply calculates the test statistic or parameters using the expected rank vector. As pointed out by Hughes and Millard (1988), the expected rank vector method furnishes the justification for employing the commonly accepted technique of splitting ranks when there are tied data (see Section 23.3.2). However, the conditional test statistic calculated using expected ranks does not have the same null distribution as in the case where there are no ties. In particular, the variance of the test statistic is smaller when ties are present (Lehmann, 1975). Consequently, a variance correction is usually required for test statistics when dealing with multiple censored data and data having ties.

Hughes and Millard (1988) present formulae for calculating expected rank vectors when there are two or more censoring levels. Additionally, they explain how to calculate the statistic required in the Mann-Kendall trend test of Section 23.3.2 and how to determine the expected value and variance of the test statistic. More specifically, for the case of nonseasonal data, one can use [23.3.1] or [23.3.4] to calculate the Mann-Kendall test statistic S or τ, respectively, using the expected rank vector. As would be expected, the Mann-Kendall test statistic is asymptotically normally distributed. Assuming that the method of censoring is independent of time, one can calculate the expected value and variance of S or τ. The expected value of S when there is multiple censoring is a function of the true value of τ, the sample size n and pattern of censoring. This expected value is determined as

$$E(S) = \tau[n(n-1)/2 - \sum_{j=1}^{p} t_j(t_j - 1)/2] \qquad [23.3.45]$$

where p is the number of tied groups in the data set and t_j is the number of data points in the jth tied group. Usually, p is the same as the number of censoring levels while the number of observations t_j in the jth tied group is the same as the corresponding number of censored observations. One can employ [23.3.2] to determine the variance of S. By computing expected ranks separately within each season, one can utilize the expected rank vector method with the seasonal Mann-Kendall test.

In some applications, the censoring levels may not be independent of time. For example, censoring levels may decrease over time due to better laboratory methods. Hughes and Millard (1988) explain how simulation can be used to determine the approximate distribution of the test statistic for this situation. Further research is still required in order to obtain theoretical results.

23.4 POWER COMPARISONS OF PARAMETRIC AND NONPARAMETRIC TREND TESTS

23.4.1 Introduction

The objective of this section is to employ *Monte Carlo experiments* to compare the powers of a specific parametric and nonparametric test for detecting trends. In particular, the ACF at lag one given in [2.5.4] and Kendall's tau in [23.3.5] constitute the parametric and nonparametric tests, respectively, which are utilized in the simulation studies. Following a brief review of these two statistics in the next two subsections, the six models that are used for generating data containing trends are described. In Section 23.4.5, the abilities in terms of power of the ACF at lag one and Kendall's tau for detecting trends are rigorously compared. Simulation experiments demonstrate that the ACF at lag one is more powerful than Kendall's tau for discovering purely stochastic trends. On the other hand, Kendall's tau is more powerful when deterministic trends are present. The results of these experiments were originally presented by Hipel et al. (1986).

23.4.2 Autocorrelation Function at Lag One

Although the ACF test at lag one could perhaps be considered to be a nonparametric test, it could also be thought of as a parametric test since according to the Yule-Walker equations in [3.2.12] the ACF at lag one is the same as the AR parameter in an AR(1) process. Nevertheless, it is presented here, because, like the nonparametric tests described in Section 23.3, it is only used for discovering the presence of trends. Unlike the intervention model, for example, the ACF test is not designed for modelling the shapes and magnitudes of trends caused by known interventions.

The theoretical definition for the ACF at lag k is given in [2.5.4] while the formula, r_k, for estimating the ACF at lag k is presented in [2.5.9]. In Section 22.3.6, the ACF at lag k is suggested as an exploratory data analysis tool and the statistical properties of r_k are discussed. Of particular interest in this section is the *ACF at lag one,* denoted by r_1, which can be used for significance testing in trend detection at the confirmatory data analysis stage. The ACF at lag one is often referred to as the *serial correlation coefficient at lag one* or the *first serial correlation coefficient.*

The estimate for the *ACF at lag k* for an evenly spaced annual series, x_t, $t = 1,2, \ldots, n$, can be calculated using [2.5.9] (Jenkins and Watts, 1968) as

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad k > 0 \qquad [23.4.1]$$

where $\bar{x}$ is the estimated mean of the x_t series. When $k = 1$ in [23.4.1], one obtains the value of r_1. For samples as small as 10, Cox (1966) observed that r_1 has an approximate normal null distribution for both normal and some nonnormal parent distributions. Knoke (1977) found empirically that the normal distribution provides an adequate approximation for determining the critical regions (the subset of the sample space for which the null hypothesis is rejected if the data fall there). The asymptotic distribution of r_1 was established by Wald and Wolfowitz (1943) and Noether (1950). In this section, critical regions for r_1 are determined by the normal approximations with the following moments (Kendall et al., 1983; Dufour and Roy, 1985)

$$mean = -1/n$$

and

$$variance = \frac{(n-2)^2}{n^2(n-1)} \qquad [23.4.2]$$

Knoke (1975) noted that r_1 is a powerful test for detecting nonrandomness for first order autoregression alternatives and that it performs reasonably well for a wider class of alternatives including the first order moving average model.

When dealing with seasonal data, a separate ACF can be estimated for each season. In Section 14.3.2, this is referred to as the *periodic ACF*. Let x_{ij} be a time series value for the ith year and jth season where there are n years of data for each of the m seasons. As in [14.3.4], for the jth season, the ACF is estimated using

$$r_k^{(j)} = \frac{\sum_{i=1}^{n-k}(x_{ij} - x_{.j})(x_{ij-k} - x_{.j-k})}{\left[\sum_{i=1}^{n}(x_{ij} - x_{.j})^2\right]\left[\sum_{i=1}^{n}(x_{ij-k} - x_{.j-k})^2\right]}, \quad k = 1,2,\cdots \qquad [23.4.3]$$

where $x_{.j}$ is the mean of the jth season.

The ACF for season j is asymptotically normally distributed with a mean of zero and a variance of $1/n$. To be more accurate, the formula in [23.4.2] could be used for calculating the mean and variance of $r_k^{(j)}$ for each season. To check if there is a trend in the jth season, one can calculate $r_l^{(j)}$ using [23.4.3] and then perform a significance test to ascertain if $r_l^{(j)}$ is significantly different from zero at the chosen level of significance. If $r_l^{(j)}$ were significantly large, this may indicate the presence of a trend. This kind of test can be done separately for each of the seasons in order to check for a trend in each season of the year. Note that when $r_j(1)$ is determined using [23.4.3], only the data within seasons j and $j-1$ are employed in the calculation. To produce an overall test for trends across seasons, Fisher's formula in [23.3.30] can be utilized.

23.4.3 Kendall's Tau

For the case of nonseasonal data, Kendall's tau, denoted by τ, is defined in [23.3.5] in terms of the Mann-Kendall statistic S given in [23.3.1]. When dealing with seasonal time series, one can use Kendall's tau for the gth season which is presented in [23.3.9]. However, for the purpose of this discussion, the nonseasonal version of Kendall's tau is entertained. Rather than defining τ in terms of a test statistic as in [23.3.5], an interesting interpretation is to express τ using probabilities. Specifically, for any two pairs of random variables (X_i,Y_i) and (X_j,Y_j), *Kendall's tau* is defined as the difference (Gibbons, 1971)

$$\tau = \pi_c - \pi_d \qquad [23.4.4]$$

where

$$\pi_c = Pr[(X_i < X_j)\cap(Y_i < Y_j)] + Pr[(X_i > X_j)\cap(Y_i > Y_j)]$$

and

$$\pi_d = Pr[(X_i < X_j)\cap(Y_i > Y_j)] + Pr[(X_i > X_j)\cap(Y_i < Y_j)]$$

In the case of no possibility of ties in either the X's or the Y's, the τ can be further expressed as

$$\tau = 2\pi_c - 1 = 1 - 2\pi_d$$

As in Section 23.4.2, let an evenly spaced yearly time series be denoted as x_t, $t = 1,2,\ldots,n$. For this sequence of observations, Kendall's τ is estimated by (Gibbons, 1971, 1976; Conover, 1980; Kendall, 1975; Hollander and Wolfe, 1973)

$$\tau = \frac{S}{\begin{pmatrix} n \\ 2 \end{pmatrix}} = \frac{N_c - N_d}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \qquad [23.4.5]$$

where N_c, N_d and S are given by

$$N_c = \sum_{i<j}^{n} \theta'_i$$

for which

$$\theta'_i = \begin{cases} 1, & \text{if } x_i < x_j \\ 0, & \text{otherwise} \end{cases}$$

and

$$N_d = \sum_{i<j}^{n} \delta'_i$$

for which

$$\delta'_i = \begin{cases} 1, & \text{if } x_i > x_j \\ 0, & \text{otherwise} \end{cases}$$

and

$$S = \sum_{i<j}^{n} \phi'_i$$

for which

$$\phi'_i = \begin{cases} 1, & \text{if } x_i < x_j \\ 0, & \text{if } x_i = x_j \\ -1, & \text{otherwise}. \end{cases}$$

The statistic S can also be expressed as

$$S = \begin{pmatrix} n \\ 2 \end{pmatrix} - 2N_d = N_c - N_d = \begin{pmatrix} n \\ 2 \end{pmatrix} \tau$$

Under the assumption that the x_i's are IID (identically independently distributed), the means and variances for S and τ, respectively, are given in [23.3.2] and [23.3.10], respectively. Kendall (1975) and Mann (1945) derive the exact distribution of S for $n \leq 10$, and, for samples as small as 10, show that the normal assumption is adequate. However, for use with the normal approximation, Kendall (1975) suggests a continuity correction which is the standard normal variate given in [23.3.3].

23.4.4 Alternative Generating Models

The six models defined in this section are used for simulating the nonseasonal data employed for comparing the powers of the ACF at lag 1 which is r_1 in [23.4.1] and Kendall's tau in [23.4.5]. The first three models contain only *deterministic trends* while the last three have purely *stochastic trends*. Furthermore, under the null hypothesis it is assumed that the time series x_t, $t = 1, 2, \ldots, n$, consists of IID random variables. As noted in Section 23.3.2, the Mann-Kendall statistic S, equivalently defined in both Section 23.4.3 and [23.3.1], is often used in place of τ which is defined in [23.4.4], [23.4.5] and [23.3.5] (Kendall, 1975; Hirsch et al., 1982; Hirsch and Slack, 1984; Van Belle and Hughes, 1984). In fact, τ and S are statistically equivalent.

For the case of a *purely deterministic trend component*, the time series, x_t, may be written as

$$x_t = f(t) + a_t \qquad [23.4.6]$$

where $f(t)$ is a function of time only and hence is a purely deterministic trend, while a_t is an IID sequence. On the other hand, a time series having a purely stochastic trend may be defined as

$$x_t = f(x_{t-1}, x_{t-2}, \cdots) + a_t \qquad [23.4.7]$$

where $f(x_{t-1}, x_{t-2}, ...)$ is a function of the past data and a_t is an innovation series assumed to be IID and with the property

$$E[a_t \cdot x_{t-k}] = 0, \quad k = 1, 2, \cdots . \qquad [23.4.8]$$

In actual practice, it may be difficult to distinguish between deterministic and stochastic trends. For example, the series plotted in Figure 23.4.1 was simulated from the model

$$(1 - B)^3 x_t = a_t \qquad [23.4.9]$$

where B is the backward shift operator, $a_t \approx NID(0,1)$ and the starting values are $x_1 = 100$, $x_2 = 101$ and $x_3 = 102$. The model in [23.4.9] is an ARIMA(0,3,0) model and the procedure for simulating with any type of ARIMA model is described in detail in Chapter 9. Based upon the shape of the graph in Figure 23.4.1, the series could probably be adequately described using a purely deterministic trend even though the correct model is purely stochastic. The same comments are also valid for the simulated sequences in Figures 4.2.1, 4.2.2, 4.3.4, and 4.3.5 of Chapter 4. Moreover, a discussion regarding deterministic and stochastic trends is provided in Section 4.6.

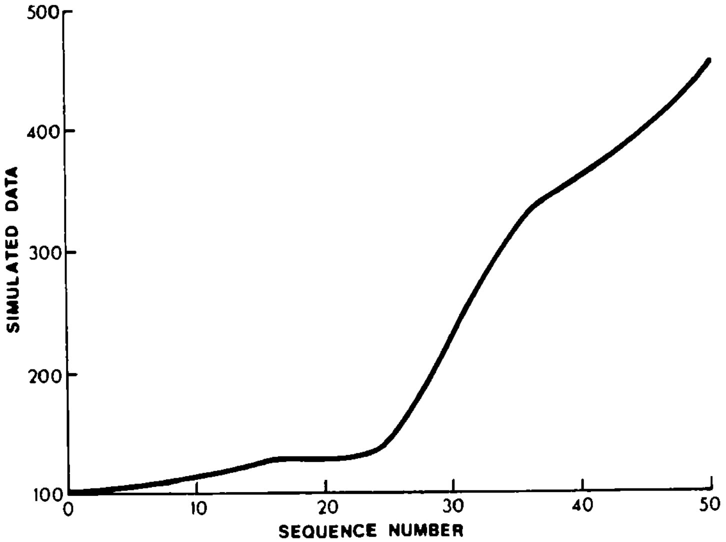


Figure 23.4.1. Simulated sequence from an ARIMA(0,3,0) model.

Box and Jenkins (1976) suggest that for forecasting purposes it is usually better to use a purely stochastic trend model provided that such a model appears to be reasonable a priori for fitting to a given time series and also provides an adequate fit. However, in water quality studies it is often of interest to test if the level of the series has changed in some way and in this case a model with a possible deterministic trend component may seem more suitable beforehand. Three deterministic models, followed by three purely stochastic models are now defined.

Linear Model

In the water resources literature, using linear regression models as alternative hypotheses is quite common (Lettenmaier, 1976; Hirsch et al., 1982; Hirsch and Slack, 1984; van Belle and Hughes, 1984). Assume x_t is given by the linear model which is also written in [4.5.2] as

$$x_t = c + bt + a_t\ ,\ t = 1,2,\ldots,n \qquad [23.4.10]$$

where $a_t \approx NID(0,\sigma_a^2)$, and c and b are constants. Without loss of generality, let $c = 0.0$.

Logistic Model

Because it is possible for a series to change rapidly at the start and then gradually approach a limit, a *logistic model* constitutes a reasonable choice for an alternative model. This model is defined as (Cleary and Levenbach, 1982)

$$x_t = M/[1 - c\left\{\exp(-bt)\right\}] + a_t\ ,\ t = 1,2,\ldots,n \qquad [23.4.11]$$

where $a_t \approx NID(0,1)$, M is the limit of x_t as t tends to infinity, and b and c are constants.

Step Function Model

Following [4.5.1], the *step function model* is defined as

$$x_t = \begin{cases} a_t, & \text{if } 0 \le t \le n/2 \\ \\ c + a_t, & \text{if } n/2 < t \le n \end{cases} \qquad [23.4.12]$$

where $a_t \approx NID(0,\sigma_a^2)$ and c is the average change in the level of the series after time $t = n/2$. The step function model is a special type of intervention model. The unit step function is defined in [19.2.3] while an intervention model that can handle step interventions is given in [19.2.9]. Besides Chapter 19, applications of intervention models to water quality and quantity time series are presented in Section 22.4.

Barnard's Model

The *Barnard model* is defined as (Barnard, 1959)

$$x_t = x_{t-1} + \sum_{i=1}^{n_t} \delta_i + a_t\ ,\ t = 1,2,\ldots,n \qquad [23.4.13]$$

where n_t follows a Poisson distribution with parameter λ, $\delta_i \approx NID(0,\sigma_i^2)$ and $a_t \approx NID(0,1)$.

Without loss of generality, let $x_1 = a_1$. Barnard (1959) developed this model for the use in quality control where there may be a series of n_t correctional jumps between measurements.

Second Order Autoregressive Model

From [3.2.4] or [3.2.5], an *AR(2) model* may be written as

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + a_t \ , \ t = 1,2, \ldots, n \qquad [23.4.14]$$

where $a_t \approx NID(0,\sigma_a^2)$. For the simulation studies executed in Section 23.4.5, $\sigma_a^2 = 1.0$ and $E(x_t) = 0.0$. In Section 3.2.2, the general expression for the theoretical ACF of an AR(p) process is given in [3.2.10] and the approach for solving for the theoretical ACF using the Yule-Walker equations in [3.2.12] is explained. Because of the correlation structure present in an AR(p) model or the AR(2) model in [23.4.14], simulated data from an AR model will contain stochastic trends. In the graph of a series simulated using an AR model such as the one in [23.4.14], a sequence of high values will often be grouped together and low values will often follow other small data points.

Threshold Autoregressive Model

The development of the *threshold autoregressive (TAR) model* is due to Tong (1977, 1978, 1983), Tong and Lim (1980) and Tong et al. (1985). Tong (1983), Tong and Lim (1980), and Tong et al. (1985) found TAR models to be suitable for modelling and forecasting riverflows.

The particular model considered here (Tong, 1983; Tong et al., 1985) is given by

$$x_t = \begin{cases} 1.79 + 0.76x_{t-1} - 0.05x_{t-2} + a_t^{(1)}, & \text{if } J_t < -1 \\ \\ 0.87 + 1.3x_{t-1} - 0.71x_{t-2} + 0.34x_{t-3} + a_t^{(2)}, & \text{otherwise} \end{cases} \qquad [23.4.15]$$

where x_t is the volume of riverflow in cubic metres per second per day, J_t is the temperature in degrees centigrade, and $a_t^{(1)} \approx NID(0,0.69)$ and $a_t^{(2)} \approx NID(0,7.18)$. The above model was estimated for the Vatnsdalsa River in Iceland for the period from 1972 to 1974.

23.4.5 Simulation Experiments

The procedures and algorithms for simulating with ARMA and ARIMA models are described in detail in Chapter 9. When a *trend* component is present, which is the case for the first three models of Section 23.4.4, the noise component is simulated separately and the deterministic component at each point in time is added to this. Note that for the first three models of Section 23.4.4, the noise component is white (i.e., it is independently distributed). Because of the white noise component, the first three models are deemed to have purely deterministic trends. However, if the noise component were correlated and were, for example, an ARMA model, a model containing a deterministic trend component plus the correlated or stochastic noise component would no longer possess purely deterministic trends. This is because the AR and MA components of the ARMA model would create a *stochastic trend* component and when this is added to the deterministic trend part of the model, the overall result would be a *mixed-deterministic-stochastic trend*. In fact, as already pointed out in Section 23.1, the general intervention model in [19.5.8] contains both deterministic trend components (i.e., the intervention terms) and a stochastic trend component (i.e., the correlated noise term). In order to be able to clearly discriminate between the powers of r_1 and Kendall's tau for detecting trends, only purely deterministic trends (the first three models in Section 23.4.4) and purely stochastic trends (the last three models in Section 23.4.4) are entertained in the simulation experiments of this section.

For each of the six generating models of the previous section, sample sizes of length 10, 20, 50 and 100 are considered. For each sample size or length of series, 1000 sequences of the same length are simulated. The null hypothesis is that each replication of a given length is IID while the alternative hypothesis is the replication contains a deterministic or stochastic trend component. For both the ACF at lag one, r_1, in [23.4.1] and Kendall's tau, τ, in [23.3.4] or [23.4.5] and a specified sample size, power functions are estimated for a significance level of 5% by the proportions of rejection of the null hypothesis from 1000 replications. As explained in Section 23.2.2 and Table 23.2.1, the proportion of rejections can be interpreted as the probability of accepting the alternative hypothesis which is the *power*.

For the estimated significance level, the test is said to be *conservative* if the estimated level is clearly less than the nominal level (in this case 0.05). On the other hand, if the estimated level is clearly greater than the 0.05, the test is said to be *optimistic*. Otherwise, the test is said to be *adequately approximated*.

Empirical significance levels and powers are given in Tables 23.4.1 to 23.4.6 for the six models defined in Section 23.4.4, respectively. Notice that except for the TAR model, for each model a range of values is used for each of the parameters and the estimated powers are given for τ and r_1 for sample sizes or series having lengths of 10, 20, 50 and 100. The standard error of any entry in the tables is $\sqrt{\pi(1-\pi)/N}$ (Cochran, 1977), where N is the number of replications and π is the true rejection rate. For example, for the estimated significance level of 5%, the standard error is $\left[\frac{0.05(1-0.05)}{1000}\right]^{1/2} = 0.0069$. The entries in the tables suggest that the critical regions are adequately determined by the null approximate distribution. The results of the simulation experiments are discussed separately for each model.

Linear Model

The findings of the simulation study for the linear model in [23.4.10] are presented in Table 23.4.1. Notice that the two tests, consisting of Kendall's tau and r_1 perform better when the standard deviation, σ_a, for the white noise term, a_t, is smaller. This implies that the better the fit of a linear regression to a time series, the greater the chance of detection of nonrandomness. For instance, for samples as small as 10, the tests are very powerful for small standard deviations. An encouraging aspect of this model is that both tests attain asymptotic efficiency quite rapidly. For example, there is considerable improvement in the power functions from $n = 10$ to $n = 20$. A noteworthy point is that τ is generally more powerful compared to r_1, even though the difference is almost negligible for $n = 50$ and $n = 100$ when both tests approach asymptotic efficiency.

Table 23.4.1. Power comparisons for the linear models with an empirical rejection rate at the 5 percent level of significance.

Parameter Values		n							
		10		20		50		100	
b	σ_a	τ	r_1	τ	r_1	τ	r_1	τ	r_1
0.00	0.05	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.01	0.05	0.335	0.138	0.995	0.749	1.000	1.000	1.000	1.000
0.01	0.50	0.057	0.039	0.072	0.043	0.487	0.080	1.000	0.690
0.01	1.00	0.053	0.040	0.050	0.043	0.146	0.042	0.769	0.131
0.01	2.00	0.050	0.041	0.036	0.047	0.073	0.042	0.268	0.065
0.05	0.05	1.000	0.990	1.000	1.000	1.000	1.000	1.000	1.000
0.05	0.50	0.121	0.063	0.632	0.194	1.000	1.000	1.000	1.000
0.05	1.00	0.066	0.049	0.208	0.064	1.000	0.655	1.000	1.000
0.05	2.00	0.050	0.040	0.084	0.043	0.669	0.121	1.000	0.927
0.10	0.05	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	0.50	0.335	0.138	0.995	0.749	1.000	1.000	1.000	1.000
0.10	1.00	0.121	0.063	0.632	0.194	1.000	1.000	1.000	1.000
0.10	2.00	0.066	0.049	0.208	0.064	1.000	0.655	1.000	1.000

Logistic Model

The results for the logistic model in [23.4.11] are given in Table 23.4.2. As is the case for the linear model, the tests perform better for the logistic model when there is a good fit, indicating nonrandomness. When $M < 1.0$, obviously the standard deviation of 1.0 used in the simulation studies tends to have a greater impact on the simulated data than the other parameters of the model. Hence, a substantial component of x_t is determined by a_t, which is random. For $M \geq 1.0$, the two tests (especially τ) prove effective for detecting the presence of trends. Finally, it can be seen that τ is more powerful than r_1, especially for cases where the logistic model describes the data fairly well ($M \geq 1.0$). There is, however, not much difference between the two tests when $n = 100$.

Step Function Model

As can be seen in Table 23.4.3, the output for the step function model in [23.4.12] indicates greater power for relatively small standard deviations (and hence fairly good fits). What is remarkable about this model is the great power of both tests even for samples as small as 10. For example, for $c = 5$, the power is at least 50% for all sample sizes. The power functions also improve as c increases. Both tests are very effective in detecting trends for even a slight shift of 0.5 in the mean level of the series. For a change of 5 in the mean level, both tests are very powerful. Even though τ is more powerful than r_1, both tests are almost equally powerful for $n \geq 50$.

Table 23.4.2. Power comparisons for the logistic models.

Parameter Values			n							
			10		20		50		100	
b	c	M	τ	r_1	τ	r_1	τ	r_1	τ	r_1
0.01	0.01	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.01	0.01	0.1	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.01	0.01	1.0	0.052	0.041	0.035	0.044	0.041	0.045	0.047	0.050
0.01	0.01	5.0	0.052	0.041	0.037	0.045	0.041	0.045	0.047	0.050
0.01	0.50	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.01	0.50	0.1	0.053	0.042	0.037	0.047	0.043	0.047	0.050	0.048
0.01	0.50	1.0	0.051	0.042	0.049	0.040	0.166	0.057	0.514	0.075
0.01	0.50	5.0	0.096	0.046	0.375	0.078	1.000	0.771	1.000	0.999
0.01	0.90	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.01	0.90	0.1	0.057	0.045	0.072	0.045	0.174	0.060	0.279	0.057
0.01	0.90	1.0	0.825	0.431	1.000	0.950	1.000	1.000	1.000	1.000
0.01	0.90	5.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	0.01	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.10	0.01	0.1	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.10	0.01	1.0	0.053	0.041	0.037	0.047	0.040	0.044	0.047	0.050
0.10	0.01	5.0	0.053	0.041	0.037	0.047	0.043	0.045	0.046	0.050
0.10	0.50	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.10	0.50	0.1	0.054	0.046	0.040	0.047	0.045	0.048	0.047	0.049
0.10	0.50	1.0	0.076	0.044	0.092	0.045	0.155	0.063	0.153	0.053
0.10	0.50	5.0	0.622	0.272	0.934	0.579	0.983	0.893	0.951	0.905
0.10	0.90	0.0	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.10	0.90	0.1	0.052	0.045	0.047	0.045	0.058	0.047	0.053	0.045
0.10	0.90	1.0	0.603	0.239	0.745	0.366	0.733	0.490	0.583	0.464
0.10	0.90	5.0	1.000	0.976	1.000	1.000	1.000	1.000	0.999	1.000

Barnard's Model

The results in Table 23.4.4 for the model in [23.4.13], are very consistent and easily comprehensible. For all sample sizes and all combinations of lambda (λ) and standard deviation (σ_a), r_1 has greater power than τ. For n as small as 50, r_1 attains asymptotic efficiency, while τ is only about 80% efficient. The power of the two tests can be well appreciated by considering the results for $n = 10$. While the power of τ is about 50% that of r_1 is always greater than 50%.

Second Order Autoregressive Model

From Table 23.4.5, one can see that the findings for the AR(2) model in [23.4.14] parallel fairly closely those of Barnard's model. The main difference between the two models is that the results here are not as dramatic as in Table 23.4.4. Here too, r_1 is more powerful than τ. As n increases, the power of r_1 increases faster than that of τ. For $n = 100$, r_1 attains almost 100% of efficiency while τ performs fairly poorly in some cases. For example, for $\phi_1 = -0.2$ and $\phi_2 = 0.5$

Table 23.4.3. Power comparisons for the step function models.

Parameter Values		n							
		10		20		50		100	
c	σ_a	τ	r_1	τ	r_1	τ	r_1	τ	r_1
0.00	0.05	0.052	0.041	0.035	0.044	0.040	0.045	0.048	0.051
0.05	0.05	0.199	0.121	0.382	0.152	0.803	0.281	0.983	0.520
0.05	0.50	0.058	0.040	0.038	0.048	0.051	0.039	0.069	0.051
0.05	1.00	0.053	0.040	0.036	0.047	0.046	0.044	0.055	0.049
0.05	2.00	0.053	0.040	0.035	0.045	0.045	0.045	0.052	0.050
0.50	0.05	0.711	1.000	0.994	1.000	1.000	1.000	1.000	1.000
0.50	0.50	0.199	0.121	0.382	0.152	0.803	0.281	0.983	0.520
0.50	1.00	0.090	0.057	0.131	0.064	0.283	0.070	0.525	0.113
0.50	2.00	0.055	0.048	0.060	0.042	0.103	0.043	0.160	0.059
1.00	0.05	0.711	1.000	0.994	1.000	1.000	1.000	1.000	1.000
1.00	0.50	0.490	0.352	0.887	0.594	1.000	0.958	1.000	1.000
1.00	1.00	0.199	0.121	0.382	0.152	0.803	0.281	0.983	0.520
1.00	2.00	0.090	0.057	0.131	0.064	0.283	0.070	0.525	0.113
5.00	0.05	0.711	1.000	0.994	1.000	1.000	1.000	1.000	1.000
5.00	0.50	0.711	1.000	0.994	1.000	1.000	1.000	1.000	1.000
5.00	1.00	0.711	0.943	0.944	1.000	1.000	1.000	1.000	1.000
5.00	2.00	0.596	0.507	0.960	0.821	1.000	0.998	1.000	1.000

the power of r_1 is 0.881 while that of τ is only 0.061 for $n = 100$.

Threshold Autoregressive Model

The specific TAR model used in the simulation experiments is given in [23.4.15]. From Table 23.4.6, one can see that the results for this TAR model are very similar to those for Barnard's model and the AR(2) model in Tables 23.4.4 and 23.4.5, respectively. The r_1 test is obviously more powerful than τ. Moreover, while the power of r_1 increases very rapidly with increasing n, the power of τ only makes a slow progression. Finally, the r_1 test is about 90% efficient for $n = 20$ and it attains 100% efficiency at $n = 50$. On the other hand, the power of τ is less than 50% even for $n = 100$.

23.4.6 Conclusions

The general deductions from the simulation experiments presented in Section 23.4.5 are that the nonparametric test, using τ, is more powerful for detecting trends for data generated from the first three models while the parametric test, utilizing r_1, is more powerful for discovering trends in synthetic sequences from the last three models. As noted in Section 23.4.4, the first three models contain deterministic trends and the last three have stochastic trends. Therefore, it is reasonable to conclude that τ is more powerful for detecting deterministic trends while r_1 is more powerful for discovering stochastic trends.

Table 23.4.4. Power comparisons for Barnard's models.

Parameter Values		n					
		10		20		50	
λ	σ_a	τ	r_1	τ	r_1	τ	r_1
1.0	0.05	0.484	0.579	0.681	0.933	0.809	1.000
1.0	0.50	0.459	0.579	0.667	0.944	0.819	1.000
1.0	1.00	0.482	0.571	0.680	0.944	0.817	1.000
1.0	2.00	0.486	0.579	0.655	0.955	0.800	1.000
2.0	0.05	0.477	0.573	0.688	0.937	0.802	1.000
2.0	0.50	0.575	0.586	0.690	0.946	0.789	1.000
2.0	1.00	0.467	0.586	0.683	0.954	0.798	1.000
2.0	2.00	0.460	0.571	0.669	0.958	0.784	1.000
5.0	0.05	0.485	0.560	0.689	0.935	0.812	1.000
5.0	0.50	0.469	0.562	0.670	0.937	0.802	1.000
5.0	1.00	0.491	0.577	0.680	0.948	0.802	1.000
5.0	2.00	0.478	0.591	0.674	0.952	0.783	1.000
10.0	0.05	0.488	0.565	0.690	0.938	0.796	1.000
10.0	0.50	0.501	0.628	0.677	0.961	0.802	1.000
10.0	1.00	0.472	0.603	0.665	0.946	0.790	1.000
10.0	2.00	0.480	0.586	0.665	0.960	0.796	1.000
20.0	0.05	0.473	0.569	0.689	0.941	0.801	1.000
20.0	0.50	0.501	0.612	0.658	0.949	0.804	1.000
20.0	1.00	0.498	0.593	0.654	0.950	0.811	1.000
20.0	2.00	0.506	0.607	0.666	0.946	0.819	1.000

In practice, it is advantageous to have both a sound physical and statistical understanding of the time series being analyzed. This will allow one to decide whether one should employ models possessing deterministic trends or whether one should use models having stochastic trends. For example, it may be better to describe certain kinds of water quality measurements using models having deterministic trends. On the other hand, for modelling seasonal riverflows, models having stochastic trends, such as a TAR model, may work well (Tong, 1983; Tong et al., 1985). In other cases, one may wish to use a model which possesses both deterministic and stochastic trends. As a matter of fact, most of the intervention models used in the water quality and quantity applications of Chapter 19 and Section 22.4 have components to model both deterministic and stochastic trends.

23.5 WATER QUALITY APPLICATIONS

23.5.1 Introduction

When executing a complex *environmental impact assessment study*, usually a wide variety of statistical tests are required in order to check a range of hypotheses regarding the statistical properties of the data. The main objective of this section is to clearly explain how both nonparametric and parametric tests can be employed in an optimal fashion to extract

Table 23.4.5. Power comparisons for the AR(2) models.

Parameter Values		n							
		10		20		50		100	
ϕ_1	ϕ_2	τ	r_1	τ	r_1	τ	r_1	τ	r_1
-1.40	-0.80	0.000	0.731	0.000	1.000	0.000	1.000	0.000	1.000
-0.70	-0.80	0.002	0.009	0.000	0.154	0.000	0.975	0.000	1.000
0.70	-0.80	0.031	0.170	0.009	0.485	0.002	0.991	0.000	1.000
1.40	-0.80	0.247	0.881	0.132	0.998	0.067	1.000	0.045	1.000
-1.20	-0.50	0.000	0.660	0.000	0.989	0.000	1.000	0.000	1.000
-0.60	-0.50	0.007	0.064	0.002	0.279	0.000	0.900	0.000	0.998
0.60	-0.50	0.072	0.223	0.050	0.470	0.036	0.932	0.031	1.000
1.20	-0.50	0.305	0.745	0.264	0.985	0.285	1.000	0.230	1.000
-0.80	-0.20	0.001	0.422	0.001	0.866	0.000	1.000	0.000	1.000
-0.40	-0.20	0.009	0.082	0.004	0.210	0.002	0.645	0.002	0.939
0.40	-0.20	0.082	0.137	0.091	0.267	0.089	0.684	0.103	0.951
0.80	-0.20	0.260	0.456	0.290	0.850	0.264	1.000	0.268	1.000
-0.40	0.10	0.010	0.207	0.010	0.470	0.008	0.840	0.007	0.991
0.40	0.10	0.180	0.180	0.187	0.386	0.248	0.814	0.245	0.985
0.80	0.10	0.390	0.452	0.536	0.872	0.633	1.000	0.625	1.000
-0.50	0.30	0.013	0.500	0.005	0.788	0.002	0.990	0.005	1.000
-0.30	0.30	0.024	0.239	0.023	0.466	0.023	0.776	0.028	1.000
0.30	0.30	0.193	0.138	0.297	0.306	0.335	0.701	0.348	0.934
0.50	0.30	0.309	0.243	0.401	0.585	0.524	0.975	0.527	0.999
-0.40	0.50	0.017	0.615	0.013	0.824	0.005	0.984	0.008	1.000
-0.20	0.50	0.046	0.319	0.059	0.470	0.088	0.682	0.061	0.881
0.20	0.50	0.175	0.143	0.302	0.250	0.377	0.578	0.421	0.831
0.40	0.50	0.288	0.211	0.470	0.513	0.610	0.933	0.686	0.999

Table 23.4.6. Power comparisons for the TAR model.

n							
10		20		50		100	
τ	r_1	τ	r_1	τ	r_1	τ	r_1
0.320	0.499	0.435	0.904	0.320	1.000	0.338	1.000

systematically relevant information from a set of water quality time series. In particular, the effects of industrial development at Nanticoke, Ontario, upon the nearshore Lake Erie water chemistry are examined in a comprehensive statistical study. This undertaking was originally carried out by the authors in conjunction with Acres International Limited of Niagara Falls, Ontario, for the Ministry of the Environment in the Canadian province of Ontario. Some of the statistical findings for the Lake Erie study given in Section 23.5.2 are also presented by Hipel et al. (1988).

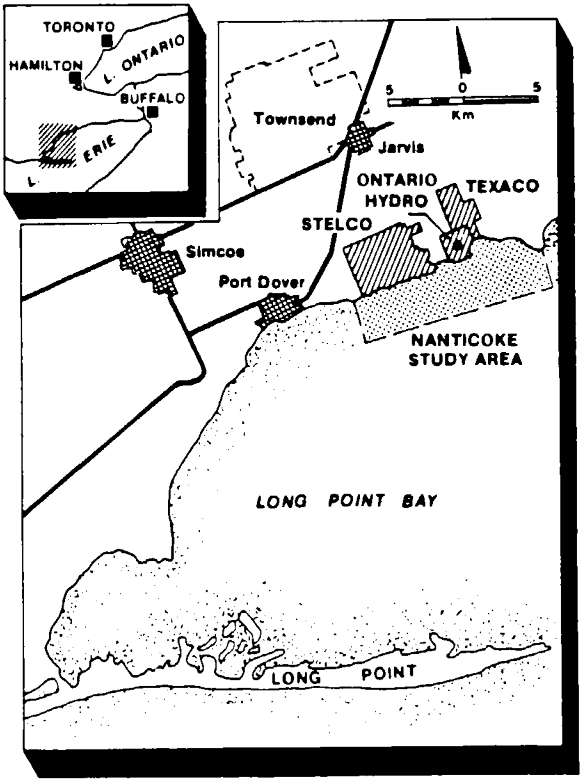


Figure 23.5.1. Location of the Lake Erie water quality study.

A map of the Long Point Bay region of Lake Erie which contains the Nanticoke study area is displayed in Figure 23.5.1. Notice that Nanticoke is situated on the north shore of Long Point Bay in Lake Erie. From the late 1960's and onwards, major industrial development took place at the town of Nanticoke. In January, 1972, Ontario Hydro's (the provincial company that generates almost all of the electrical power in Ontario) 4000 MW fossil-fueled thermal generating station commenced operations. Texaco Canada Inc. constructed an oil refinery which came into operation in November, 1978. Finally, the Steel Company of Canada (Stelco) built a steel mill that started to produce steel in May, 1980. Because the foregoing industrial projects could adversely affect the water quality of Lake Erie, the goal of the Nanticoke study was to detect trends in various water quality variables.

Water quality monitoring began in 1969 at eight (stations 112, 501, 648, 518, 810, 1008, 1016, and 994) of the fifteen sampling stations shown in Figure 23.5.2. Since 1969, the remaining seven of the fifteen stations were added to the network. Many of the stations were sampled at more than one depth, such as near the surface and near the bottom. Unfortunately, as is also the case for the water quality data examined in Chapter 22 and Section 24.3.2, the water quality series measured in Lake Erie contain observations separated by unequal time intervals, many of

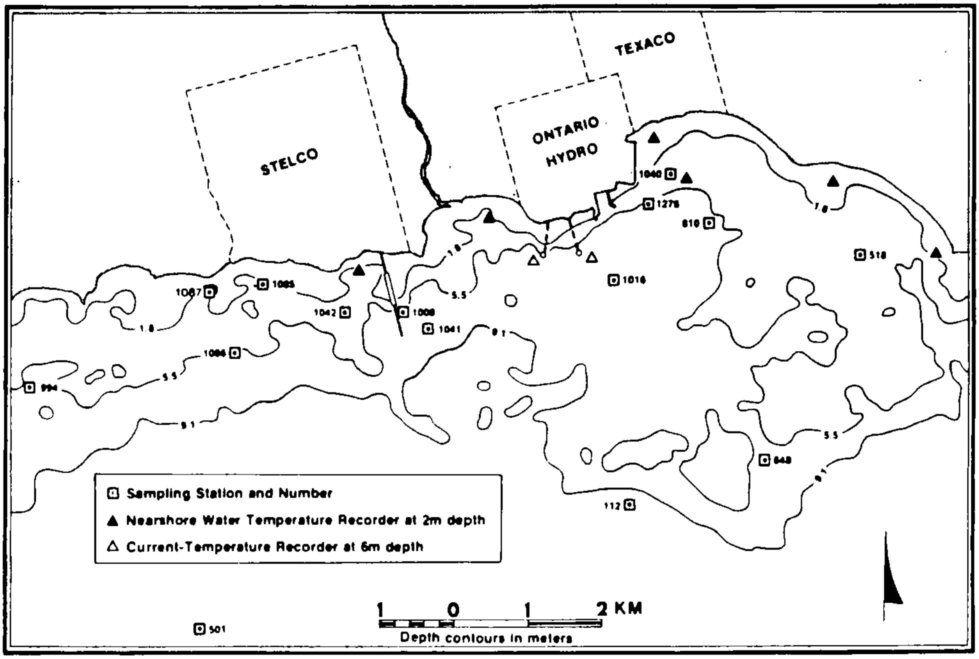


Figure 2. Sampling Stations at Long Point Bay in Lake Erie.

Figure 23.5.2. Sampling stations at Nanticoke in Long Point Bay, Lake Erie.

which are relatively long. Although not all of the water quality variables were assessed for trends, about 50 different water quality variables were sampled across all of the stations and the more important variables were collected at each of the stations. Detailed statistical testing for the presence of trends and other statistical characteristics were carried out for 14 water quality variables (see Table 23.5.6 for a list of these 14 variables) measured at Stations 501, 810, 994, 1085 and 1086. In the next section, the statistical procedures used in the study are outlined and some representative results are given. Finally, the *trend analysis methodology* put forward in the upcoming section constitutes one of the three overall trend procedures described in the book and summarized in Table 1.6.1.

23.5.2 Trend Analysis of the Lake Erie Water Quality Series

Selecting Appropriate Statistical Tests

To detect trends and uncover other statistical properties of the Lake Erie water quality time series, appropriate statistical tests must be employed. In order to have the highest probability of discovering suspected statistical characteristics which may be present in the time series, one must select the set of tests that possess the best capabilities for uncovering the specified statistical properties. To accomplish this, *one must be cognizant of both the general statistical properties of the data and the main attributes of the statistical tests*. For example, with respect to the characteristics of the data, one should be aware of properties such as the quantity of data, large time gaps where no measurements were taken, outliers and data which fall below the *detection limits* (see Section 23.3.2 for discussions about detection limits in water quality time series). As discussed in Sections 19.2.3 and 22.3, often these properties of the data are known in advance or else are revealed using *exploratory data analysis* tools. From the point of view of a statistical test that can be used, one should know key facts which include the specific null and alternative hypotheses that the given statistical test is designed to check, the major distributional assumptions underlying the test, the types of samples with which the test can be used, and the kinds of measurements that can be utilized with the test. By being cognizant of the properties of the data and the main capabilities of a wide range of both *parametric and nonparametric tests*, one can choose the most appropriate statistical tests for testing specified hypotheses such as the presence of trends. As mentioned in Section 23.3.1, summaries and charts regarding the capabilities of both nonparametric and parametric tests are available in many well known statistical texts. The handbook of Sachs (1982), for instance, is very helpful for locating the most appropriate parametric and nonparametric tests to employ in a given study. Table 23.1.1 provides a list of the nonparametric tests described in Chapter 23. The tests that are eventually selected can then be used at the *confirmatory data analysis* stage for hypothesis testing (see Section 23.2 for a general discussion of statistical tests).

The particular statistical methods used in the Lake Erie study are listed in Table 23.5.1. Notice that for each statistical method, the general purpose of the technique is described and the specific reason for using it in the Lake Erie study is explained. Furthermore, if the method is described in the text, the location is cited. Otherwise, an appropriate reference is given. The first four statistical methods in Table 23.5.1 constitute exploratory data analysis tools while the remaining methods are usually employed at the confirmatory data analysis stage. The nonparametric tests given in the table are marked with asterisks. Notice that the last statistical method, regression analysis, is discussed in detail in Section 24.2.3. All of the statistical methods listed in Table 23.5.1 were applied to each of the 14 specified water quality variables at each of the 5 stations, consisting of Stations 501, 810, 994, 1085 and 1086. To clearly explain how an environmental impact assessment project is carried out, some of the informative results of the Lake Erie study are now presented for the methods marked with a cross in Table 23.5.1. For explaining how the techniques are used in practice, the chloride water quality (mg/*l*) and total phosphorous (mg/*l*) variables are used the most. Finally, for a description of water quality processes, the reader can refer to the book of Waite (1984), as well as other authors.

Data Listing

For the chloride variable at Station 501 in Figure 23.5.2, 173 observations are available in mg/*l* from July 13, 1970, to November 19, 1979. The first 25 of these measurements are listed in Table 23.5.2. To calculate the day number, the start of January 1, 1969, is taken as day number 0.0. The component to the right of the decimal for a day number refers to the fraction of a 24 hour period at which the measurement was taken. Consequently, from Table 23.5.2, the first observation was taken on July 13, 1970, at 16 hours and 40 minutes, which has the day number 558.611. The gap, expressed in number of days between adjacent observations, is given

Table 23.5.1 Statistical methods used in the Lake Erie water quality study.

METHOD	GENERAL PURPOSE	SPECIFIC PURPOSE in the Study
Data listing†	For each series, want to know exact values, dates of measurements, depths of measurements, and station number (Section 23.5.2, Table 23.5.2).	Same as under general purpose.
Graphs of the data†	Visually detect main statistical characteristics of a series (Section 22.3.2).	Visually detect trends and see spacing of the observations.
Tukey five number summary†	Describe how the observations in a series are distributed in each season (Section 23.3.3).	Describe how measurements in a water quality series are distributed in each month.
Box and whisker graphs†	See a graphical display of the five number summary for each season in a series (Section 23.3.3).	See a plot of the five numbers for each month in a series.
Seasonal Mann-Kendall test*†	Check for trends in a series for each season of the year (Section 23.3.2).	Check for trends in a series for each month.
Fisher's combination test†	Combining tests of hypotheses (Section 23.3.4).	Test for a trend across all the months in a series by combining the significance levels from the seasonal Mann-Kendall tests for each month into a χ^2 statistic (see [23.3.30]).
Wilcoxon signed rank test*†	Determine whether the medians of two samples are the same (Appendix A23.2).	Find out whether or not measurements taken at two different depths at exactly the same time possess the same median.
Confidence interval for the median*†	Calculate a confidence interval for the difference in the medians between two samples (Appendix A23.2).	Calculate 95% confidence interval for the difference in medians between measurements taken at two different depths at exactly the same time. If zero is not contained in the confidence interval, the two medians are significantly different from one another.
Kendall rank correlation*	Determine whether or not two series are independent of one another (Appendix A23.1).	Ascertain if measurements taken at the same time at alternate depths are correlated with one another.
Cross correlation function†	Determine whether or not two series are independent of one another (see Section 22.3.4 and also Kendall (1975, Ch. 2)).	Find out if measurements taken at the same time at alternate depths are correlated with one another.
Pitman's test for equality of correlated variance	Ascertain whether or not two correlated variances are the same (Pitman, 1939).	Determine if the variances of samples taken at two different depths are the same.
One way analysis of variance	Determine if the means across k samples are significantly different from one another (Sachs, 1984, pp. 501-509). It is assumed that the k populations are normally independently distributed and have equal variances.	Find out whether or not the means among replicated samples are the same.
Kruskal-Wallis test*†	Nonparametric test to check whether or not the distributions or means across k samples are the same (Appendix A23.3). The observations are assumed to be independent of one another and follow the same distribution.	Determine if the means among replicated samples are the same.
Regression analysis†	Parametrically model relationships within a series and among series.	Determine the best data transformation, ascertain the components required in a regression model, and estimate both the average monthly and annual values for a series (see Section 24.2.3).

* - nonparametric test

† - applications for this method are given in the text

in the third column. Each measured value in mg/*l* along with the depth in meters at which the sample was taken are presented in the fourth and fifth columns, respectively. Notice that the extreme values consisting of the outside and far-out values defined in Section 22.3.3 are marked.

Table 23.5.3 shows the number of available chloride measurements by month and year for Station 501. Notice, for example, that no measurements were taken during January, February and March across all of the years. Additionally, no observations are available for the years 1969, and 1980 to 1983.

Table 23.5.2. Data listing of chloride measurements (mg/*l*) at Station 501, Long Point Bay, Lake Erie.

Day Number	Date	Gap	Measured Value	Depth (m)
558.611	July 13, 1970		27.0*	4.80
588.604	Aug. 12, 1970	30	26.0*	6.20
859.726	May 10, 1971	271	35.0**	6.10
887.632	June 7, 1971	28	25.0	6.20
915.736	July 5, 1971	28	24.0	6.70
944.635	Aug. 3, 1971	29	24.0	6.40
971.615	Aug. 30, 1971	27	23.0	6.20
999.625	Sep. 27, 1971	28	25.0	6.40
1028.604	Oct. 26, 1971	29	25.0	6.00
1197.628	Apr. 12, 1972	169	25.0	6.00
1223.802	May 8, 1972	26	25.0	6.50
1253.660	June 7, 1972	30	24.0	6.20
1280.708	July 4, 1972	27	24.0	6.50
1308.618	Aug. 1, 1972	28	24.0	6.30
1338.646	Aug. 31, 1972	30	25.0	6.00
1365.503	Sep. 27, 1972	27	24.0	6.50
1419.635	Nov. 20, 1972	54	23.0	5.50
1622.521	June 11, 1973	203	23.0	1.00
1622.521	June 11, 1973	0	24.0	10.60
1650.635	July 9, 1973	28	23.0	1.00
1650.635	July 9, 1973	0	23.0	10.60
1679.545	Aug. 7, 1973	29	23.0	1.00
1679.545	Aug. 7, 1973	0	24.0	10.60
1707.597	Sep. 4, 1973	28	24.0	1.00
1707.597	Sept. 4, 1973	0	25.0	11.00
.	.	.	.	.

Remarks:

1. January 1, 1969 at 0:00 AM is taken as Day Number 0.0.
2. Outside values are indicated by *.
3. Far-outside values are indicated by **.
4. For independent and identically distributed normal variables, the expected percentages of outside and far-outside values are 0.76% and 0.000013%, respectively.

The *Tukey 5-number summary* defined in Section 22.3.3 is also listed in Table 23.5.3. In addition, the numbers of observed and expected outside and far-out values are given. The expected number of outside and far-out values are calculated by assuming that the data follow a normal distribution (see Section 22.3.3). To calculate the expected number of outside values, one multiplies 173 (the total number of observations in Table 23.5.3) times 0.0076 (the probability of having outside values if the data are NID) to obtain the expected figure of 1.3148 shown in

Table 23.5.3. Number of chloride measurements (mg/l) at Station 501 according to the month and year.

Number of measurements available by month and year

	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	TOTAL
Jan																0
Feb																0
Mar																0
Apr				1		2	2	2	2	2	3					14
May			1	1		4	2	5	4	8	6					31
Jun			1	1	2	2		2	2	4	3					17
Jul		1	1	1	2	4	2	2	4	2	3					22
Aug		1	2	2	2	2	4	7	2	7						29
Sep			1	1	2	2	2	2	4	2	3					19
Oct			1		2	2	2	2		4	3					16
Nov				1	2		2	3	4	6	3					21
Dec						2				2						4
TOTAL	0	2	7	8	12	20	16	25	22	37	24	0	0	0	0	173

Tukey five-number summary: 19.0 20.4 21.0 22.5 35.0

	Observed	Expected
Number of Outside values	2	1.314800
Number of Far-outside values	1	0.000225

Table 23.5.3. Likewise, to determine the expected number of far-outside values one multiplies 13×10^{-8} times 173 to get 0.000225.

Graphs of the Data

Graphs of time series are presented throughout the text for a wide range of series while a discussion regarding the usefulness of graphs as exploratory data analysis tools is given in Sections 5.3.3 and 22.3.2. Figures 1.1.1 and 19.1.1, for example, displays a plot of 72 average monthly phosphorous data points (mg/l) from January, 1972, until December, 1977, for measurements taken downstream from the Guelph sewage treatment plant located on the Speed River in Ontario, Canada. In the figure, it can be seen that conventional phosphorous treatment has dramatically decreased the mean level of the series after the intervention date when the tertiary treatment was implemented. The black dots indicate locations where data are missing and hence had to be estimated. In Section 19.4.5, intervention analysis is used to model the effects of the phosphorous treatment and estimate the missing observations.

A simple approach to display effectively the statistical characteristics of a data set using ordinary output paper from a computer is to employ a *jittered one dimensional plot*. The data are plotted horizontally between the smallest observation on the left and the largest value on the right. The exact magnitudes of the smallest and largest values are given as part of the 5 number summary shown in Table 23.5.3. In the plot, the letters $a,b,c,\ldots,$ denote $1,2,3,\ldots,$ data points, respectively.

Figure 23.5.3a displays the jittered plot for all of the chloride data at Station 501. When the chloride data are plotted according to each month across all years as in Figure 23.5.3b, the manner in which the data are distributed according to each season can be observed. Notice from the jittered plot of chloride according to each year in Figure 23.5.3c, that the chloride level is

```
  b  e e d  a a    a   c  a
  b  e e d  b a a  b   c  a
a b  e e e  b b a  b  ac  a
a b  e e e  b b a aba ac  b   a   a                                     a
  b  e e e  b a a  b   c  a
  b  e e d  a a    b   c  a
  a  e d d  a a    a   c  a
     Note:  a,b,c,... etc. denote 1,2,3,... data points, respectively.
```

(a) Jittered one-dimensional data plot of measured values.

```
Jan
Feb
Mar
Apr   d a  d a   a    b      a
May   c ℓ  e a b      b aab  a                                       a
Jun     f  d c        a   b  a
Jul   b e  d b     b aba  b          a
Aug        e g c c b  d   c  a   a
Sep   b b  g b   b        b  b
Oct   b a    h b          b  a
Nov b   a  e e d a    a   a  a
Dec          b   b
```

(b) Jittered one-dimensional data plot of measured value by month.

```
1968
1969
1970                              a   a
1971                  a   b   c                                      a
1972                  a   d   c
1973                  d   f   b
1974  b b  b b   b b abaaab
1975         a d e b  d
1976    a  e ℓ e b
1977       m g b
1978  d ℓ  ℓ i
1979b g m  b
1980
1981
1982
1983
```

(c) Jittered one-dimensional data plot of measured value by year.

Figure 23.5.3. Jittered one-dimensional plots of the chloride observations (mg/*l*) at Station 501, Long Point Bay, Lake Erie.

decreasing over the years. A jittered plot of depths at which the chloride observations are taken is presented in Figure 23.5.4.

```
     Tukey five-number summary of all the depths (m) used
         1.00      1.00      1.50     11.00     13.00
               Jittered one-dimensional data plot of depths
i  e                                                          a   d  b
i  e                           aa a                     a   a a   d  b
i  e                        a  aa a               a     a   a b a d  b
i  e                    a   a  aaaa a             a  a  a   a b a da c   a
i  e                           aa a               a     a   a a a d  b
i  e                           aa a                     a     a   d  b
i  d                                                          a   d  b
     Note:  a,b,c,... etc. denote 1,2,3,... data points respectively
```

Figure 23.5.4. Jittered one-dimensional plot of depths at which the chloride observations at Station 501 in Long Point Bay, Lake Erie, are taken.

Box-and-Whisker Graphs

The box-and-whisker graph, which is based upon the 5-number summary (Tukey, 1977) is described in Section 22.3.3 where illustrative plots are also given. When entertaining seasonal data such as monthly or quarterly data, it is instructive to calculate a 5-number summary plus outside and far-outside values for each season. For the given total phosphorous data (mg/*l*) at Station 501, Figure 23.5.5 depicts box-and-whisker graphs that are commonly referred to as box plots. In this figure, the data have not been transformed using a Box-Cox transformation from [3.4.30]. The *far-out values* are indicated by a circle in Figure 23.5.5, where far-out values are not marked if there are four or less data points for a given month. Below each month is a number which gives the number of data points used to calculate the box-and-whisker graph for that . When there are not many data points used to determine a box-and-whisker plot for a given month, any peculiarities in the plot should be cautiously considered. The total number of observations across all the months is listed on the right below the x axis.

For a given month in a box-and-whisker diagram, symmetric data would cause the median to lie in the middle of the box or rectangle and the lengths of the upper and lower whiskers would be about the same. Notice in Figure 23.5.5, for the total phosphorous data at Station 501, that the whiskers are almost entirely above the rectangles for almost all of the months and there are 8 far-out values above the boxes. This lack of symmetry can at least be partially rectified by transforming the given data using the Box-Cox transformation of $\lambda = 0$ in [3.4.30]. By comparing Figure 23.5.5 to Figure 23.5.6, where natural logarithms are taken of the total phosphorous data, the improvement in symmetry can be clearly seen. Furthermore, the Box-Cox transformation has reduced the number of far-out entries from 8 in Figure 23.5.5 to 5 in Figure 23.5.6.

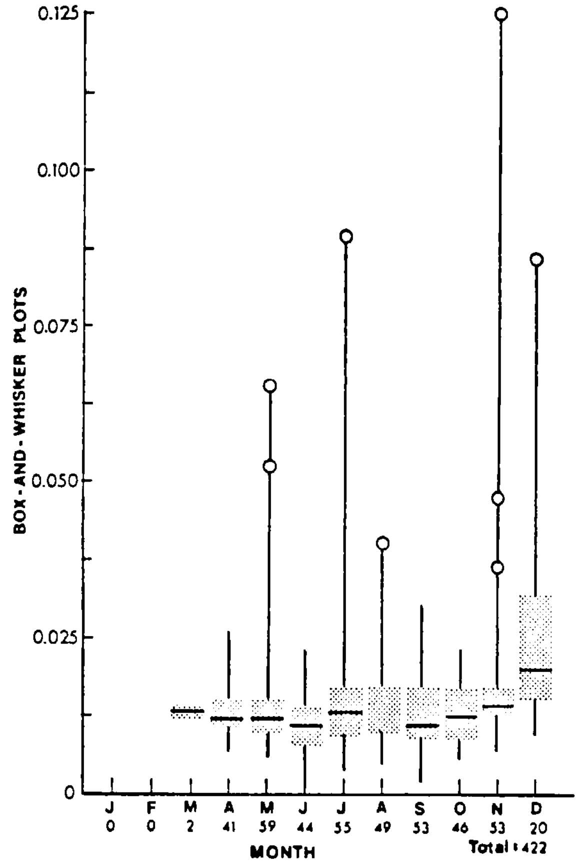


Figure 23.5.5. Box-and-whisker plots of the total phosphorous (mg/*l*) data at Station 501, Long Point Bay, Lake Erie, from April 22, 1969, to December 13, 1983.

As is also explained in Section 23.3.3, Box-and-whisker plots can be employed as an important exploratory tool in *intervention studies*. If the date of the intervention is known, box-and-whisker diagrams can be constructed for each season for the data before and after the time of the intervention. These two graphs can be compared to ascertain for which seasons the intervention has caused noticeable changes. When there are sufficient data, this type of information is crucial for designing a proper intervention model to fit to the data at the confirmatory data analysis stage (see Section 19.2.3).

For the Nanticoke data, there are two major interventions. First, Ontario Hydro built a fossil-fuelled electrical generating plant which began operating in January, 1972, and came into full operation by about January 1, 1976. Because not much data are available before 1972, January 1, 1976, is taken as the intervention data at which water quality measurements near the

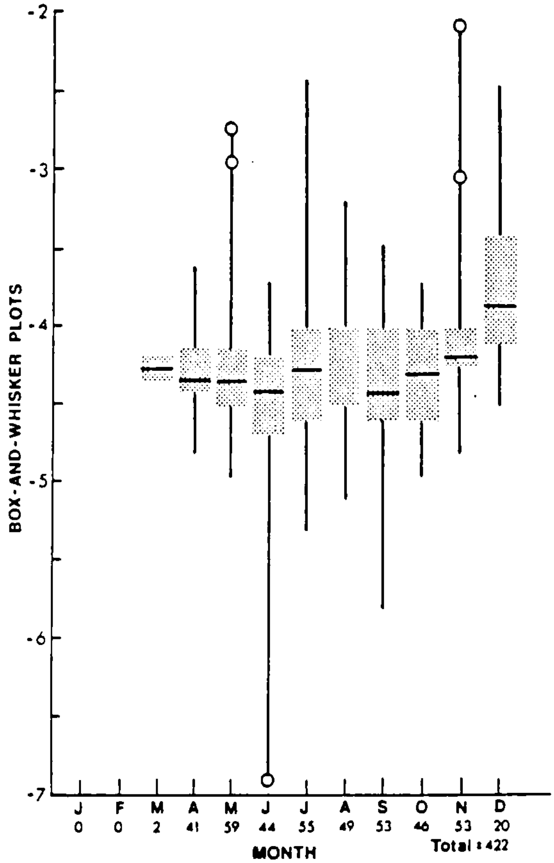


Figure 23.5.6. Box-and-whisker plots of the logarithmic total phosphorous (mg/*l*) data at Station 501, Long Point Bay, Lake Erie, from April 22, 1969, to December 13, 1983.

Ontario Hydro plant may be affected. Of the 5 stations analyzed, only Station 810 is close to the plant. For each of the water quality variables measured at Station 810, box-and-whisker plots are made before and after the intervention date in order to qualitatively discover any possible statistical impacts of the intervention.

Second, the Steel Company of Canada (Stelco) plant came into operation about April 1, 1980. Because sites 501, 994, 1085, and 1086 are relatively close to the Stelco factory, box-and-whisker graphs are made before and after the intervention for each water quality time series at each station. Figures 23.5.7 and 23.5.8 display the box-and-whisker graphs for the natural logarithms of the total phosphorous data at Station 501 before and after the Stelco intervention, respectively. The dates in brackets in the titles for Figures 23.5.7 and 23.5.8 indicate the intervals of time for which measurements were taken before and after the intervention, respectively.

When these two graphs are compared, it appears that for most of the months there is a slight drop in the median level after the intervention. Using an intervention model based upon a regression analysis design (see Chapter 24 and references therein), a confirmatory data analysis could be executed to ascertain the magnitudes of the changes in the monthly means and if they are significant. Furthermore, one should also take into account overall changes in Lake Erie by considering measurements at locations outside of the Nanticoke region.

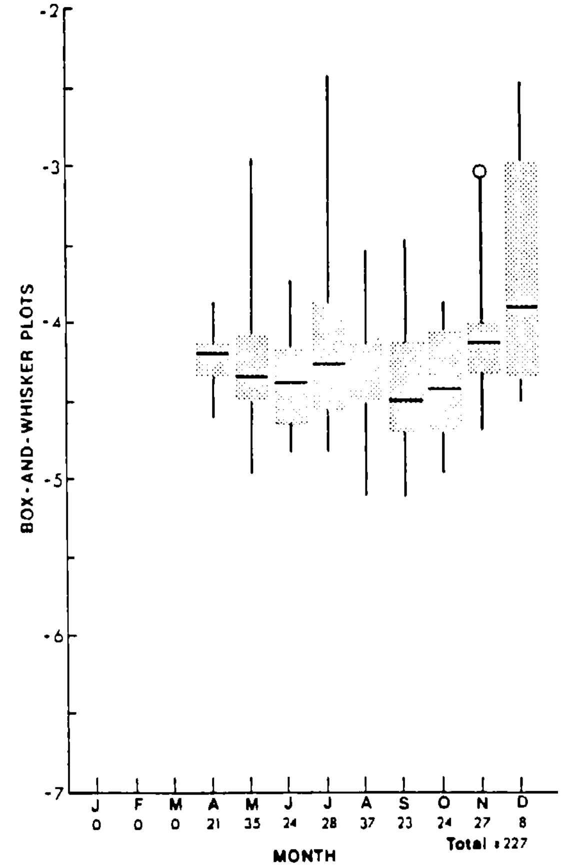


Figure 23.5.7. Box and whisker plots of the logarithmic total phosphorous data (mg/*l*) at Station 501, Long Point Bay, Lake Erie, before April 1, 1980 (data available from April 22, 1969, to November 19, 1979).

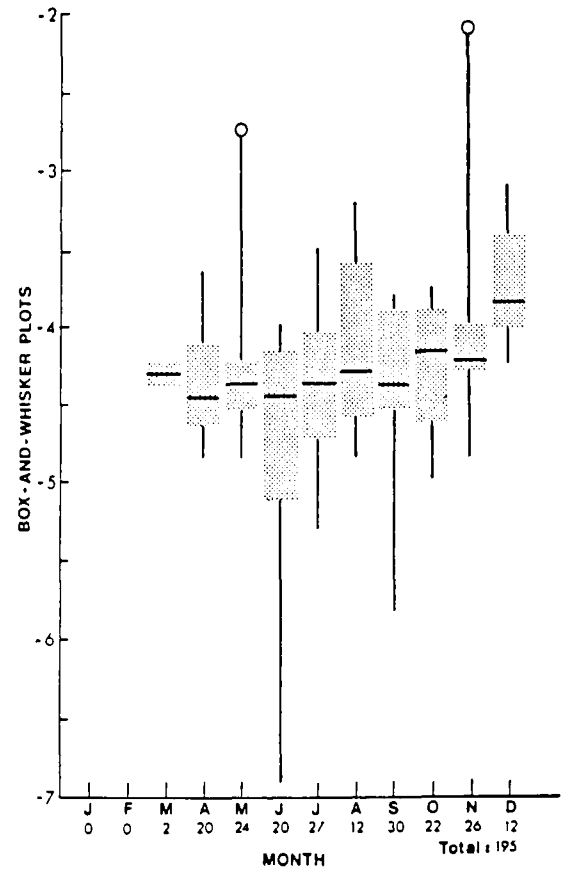


Figure 23.5.8. Box and whisker plots of the logarithmic total phosphorous (mg/*l*) data at Station 501, Long Point Bay, Lake Erie, after April 1, 1980 (data available from April 13, 1981 to December 13, 1983).

A third intervention in the Nanticoke region is due to the Texaco oil refinery which began production in November, 1978. Because the discharge from this plant is relatively quite small, the possible effects of the Texaco intervention are not considered in this study.

Seasonal Mann-Kendall Tests

For a given water quality variable at a specified station, the seasonal Mann-Kendall test can be used to detect trends in each month of the year. A detailed description of this test is given in Section 23.3.2. For the case of the chloride measurements taken at Station 501, Table 23.5.4 presents results of the seasonal Mann-Kendall and other related tests. Each entry in the table of

Table 23.5.4. Trend analysis of monthly median values using the seasonal Mann-Kendall test for the chloride series (mg/*l*) at Station 501, Long Point Bay, Lake Erie.

Monthly Median Values times 10

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1970							270	260				
1971					350	250	240	235	250	250		
1972				250	250	240	240	245	240		230	
1973						235	230	235	245	240	245	
1974				230	238	200	211	225	205	210		220
1975				215	230		225	225	220	215	215	
1976				205	205	207	210	210	210	210	215	
1977				205	205	210	205	215	205		210	
1978				197	200	200	200	205	202	210	205	210
1979				195	200	205	200		195	195	190	
tau				-0.98	-0.96	-0.62	-0.93	-0.86	-0.82	-0.82	-0.88	
SL(%)				0.39	0.17	4.61	0.03	0.22	0.33	1.87	0.98	

Combination of Scores and their Variances

Sum	Variance	SL
-196.0	5.59000×10^{2}	1.13324×10^{-16}

Fisherian Combination of the Significance Levels

CHI-SQ	DF	SL
87.01	16	0.00000

years versus months is the median value for a given month and year. By utilizing [23.3.4] or [23.3.9] for the data in a specific month across all of the years for which data are available, Kendall's tau can be determined. The observed value of S_g for each month which is calculated using [23.3.7] is not displayed in the table. Because the observed τ value for each month is negative, this indicates that there may be a decreasing trend in each month. Consider, for instance, the month of April for which the calculated τ value is -0.98. Since the SL (significance level) for this month is 0.39%, this strongly suggests that the null hypothesis of having identically independently distributed data should be rejected in favour of accepting the alternative hypothesis of there being a monotonic decreasing trend. Notice that for each month the SL is not greater than 5% and usually less than 1%. Consequently, one would expect that across all of the seasons, a combination test would confirm the presence of an overall decreasing trend. The seasonal Mann-Kendall test statistic in [23.3.11] has a magnitude of -196.0 and a very small significance level. In addition, Table 23.5.4 shows that the significance level of Fisher's combination test in [23.3.30] is also very small. Hence, both of the combination tests indicate that there is an overall trend which is decreasing due to the negative sign of S_g or τ in each season and also S' in [23.3.11] across all of the seasons. As noted earlier, this decreasing trend over the years is also readily apparent in the jittered plot of chloride according to each year in Figure 23.5.3c.

For each of the fourteen water quality variables at each of the five stations where there are sufficient data, the seasonal Mann-Kendall test in [23.3.7] or [23.3.9] is applied. Consider Table 23.5.5 which summarizes the results for chloride. Notice that at Stations 501, 810 and 994, there are obvious decreasing trends (indicated by the negative signs) for all the months for which data are available at all three sites. Except for two cases where the significance level is a = 10%, all

Table 23.5.5. Seasonal Mann-Kendall tests for trend in the chloride series (mg/*l*) for stations at Long Point Bay, Lake Erie.

MONTHS	STATIONS				
	501	810	994	1085	1086
January					
February					
March					
April	-c	-c	-b		
May	-c	-c	-d		
June	-b	-a	-a		
July	-d	-c	-c		
August	-c	-c	-c		
September	-c	-c	-c		
October	-b	-c	-c		
November	-c	-d	-d		
December					

Note:

1. a, b, c, d denote significance levels of 10, 5, 1, 0.1 percent, respectively.
2. # denotes result not significant at 10 percent.
3. Otherwise, a blank indicates insufficient data.
4. A positive or negative value of tau is indicated by + or -.

of the significance levels are 5% or less. Consequently, these trends are significant. For a given month and station, one should certainly reject the null hypothesis that the chloride data are independently and identically distributed. Table 23.5.5 shows that sufficient data for executing a seasonal Mann-Kendall test are not available for the months of January, February and March at Stations 501, 810 and 994. Also, there are not enough observations for all of the months at Stations 1085 and 1086.

Another water quality variable for which there may be decreasing monthly trends is specific conductance for Stations 501, 810 and 994. However, as is the case for chlorophyll *a*, for most of the water quality variables across most of the months and stations, significant trends are not detected by the seasonal Mann-Kendall test.

As explained earlier, along with the seasonal Mann-Kendall test, for a specified water quality variable and station one can combine the monthly results using the combined score method in [23.3.11] and the Fisherian combination in [23.3.30]. Table 23.5.6 summarizes these two types of combination results for the fourteen water quality variables across all of the stations. When interpreting these results one should keep in mind the limitations of the combination approaches described in Sections 23.3.2 to 23.3.4. Notice in Table 23.5.6 for the chloride variable that the results are highly significant for Stations 501, 810 and 994 for both combination tests (all the significance levels are $d = 0.1\%$). Consequently, there are obvious trends in chloride across the months at all three sites. Due to the negative signs in Table 23.5.5, the trends are decreasing.

Table 23.5.6. Combined score tests from [23.3.11] and Fisher's combination results using [23.3.30] for the 14 water quality variables at Long Point Bay, Lake Erie.

	STATIONS				
	501	810	994	1085	1086
turbidity (FTU)	a,#	#,#	c,b	#,#	#,#
specific conductance (μs/cm)	d,d	d,d	d,d	#,#	#,#
lab pH	#,#	#,#	#,#	#,d	#,#
chloride (mg/*l*)	d,d	d,d	d,d		
ammonia - N (mg/*l*)	d,a	c,b	b,#	#,#	#,#
morganic -N (mg/*l*)	#,#	#,#	#,#	#,d	
filtered total Kjeldahl N	c,#	#,c	a,#	b,#	#,a
Kjeldahl organic N (mg/*l*)	b,#	#,a	#,#	#,#	#,a
chlorophyll a	#,#	b,#	a,#	a,#	#,d
chlorophyll b	c,#	d,a	#,#	#,#	#,#
phytoplankton density					
filtered reactivce phosphate	d,d	b,#	d,c	#,#	#,#
total phosphorous (mg/*l*)	d,a	a,b	d,c	#,#	#,#
iron (mg/*l*)	d,b	#,a	c,b	#,#	#,#

Note:

1. a, b, c, d denote significance levels of 10, 5, 1, 0.1 percent, respectively.
2. # denotes result not significant at the 10 percent level.
3. Otherwise, a blank indicates insufficient data.
4. The combined Kendall test for trend in [23.3.11] is the first entry in each cell while the second entry is the Fisherian combination calculated using [23.3.30].

Wilcoxon Signed Rank Tests

Detailed descriptions of the Wilcoxon signed rank test along with the related confidence interval for the median are given in Appendix A23.2. The purpose of using these tests is to ascertain whether or not paired measurements taken at the same time and alternate depths are significantly different from one another.

Table 23.5.7 displays the statistical analyses of paired measurements for chloride at Station 501. At the top of the table, the availability of paired measurements by month and year is shown. Tukey 5-number summaries for both the depths and measured values at the shallow depths explain how both the depth and measured values are distributed. Likewise, for the deep samples, Tukey 5-number summaries are given for the depths of the deep measurements and the measurements themselves. Notice that the distributions of the measurements appear to be the same at the shallow and deep depths according to the Tukey 5-number summaries. When the paired measurements, denoted by X and Y for the shallow and deep depths, respectively, are subtracted from one another, the Tukey 5-number summary of $X - Y$ shows that the subtracted values are symmetrically distributed about zero. Hence, the depth of the measurement does not appear to affect the distribution of chloride.

Table 23.5.7. Analyses of paired measured chloride (mg/*l*) values at the same time and alternate depths at Station 501, Long Point Bay, Lake Erie.

Number of paired (shallow, deep) measurements available by month and year

	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	TOTAL
Jan																0
Feb																0
Mar																0
Apr						1	1	1	1	1	1					6
May						2	1	1	2	4	2					12
Jun					1	1		1	1	2						6
Jul					1	2	1	1	2	1						8
Aug					1	1	2	2	1	2						9
Sep					1	1	1	1	2	1						7
Oct					1	1	1	1		2						6
Nov					1		1		2	3						7
Dec						1				1						2
TOTAL	0	0	0	0	6	10	8	8	11	17	3	0	0	0	0	63

X: Shallow Samples

	Depths(m)	Measured Values
Extreme:	1.00	19.4
Hinge:	1.00	20.5
Median:	1.00	21.0
Hinge:	1.00	22.0
Extreme:	1.50	25.0

Y: Deep Samples

	Depths(m)	Measured Values
Extreme:	5.50	19.4
Hinge:	11.00	20.5
Median:	11.50	21.0
Hinge:	11.92	22.0
Extreme:	13.00	25.0

Five-number summary of paired X-Y: -1.0 0.0 0.0 0.0 1.0

Two-sided Wilcoxon signed rank test for paired measured values at the same time and at alternate depths:

Number of pairs is 18. Wilcoxon Statistic is 84.0. SL is 0.96526.
95% confidence interval for the median difference (X-Y) is (0.0000, 0.0000).

The Wilcoxon signed rank test can be employed as a nonparametric test to check whether or not measurements taken at two different depths at exactly the same time have the same median. As noted in Appendix A23.2, the values in a pair are not used if they are equal. As can be seen in Table 23.5.7, there are 18 pairs which do not have equal values. For these 18 pairs of chloride measurements, the value of the Wilcoxon statistic in [A23.2.3] is 84 with a SL of 0.97. Because of this very large SL, one can conclude that the medians or means of the paired chloride measurements taken at alternate depths are not significantly different from one another. Because zero is contained within the 95% confidence interval for $X - Y$, the fact that the medians are the same is further substantiated. Since the Wilcoxon test is only used when the number of paired samples having unequal measurements is greater than 7, results are not shown for some of the water quality variables at different stations.

Table 23.5.8. Summary of the Wilcoxon tests for the equality of medians of paired X and Y, where X is the shallow sample value and Y is the deep sample value.

	Stations				
	501	810	994	1085	1086
turbidity (FTU)	-c	-#	-d		-a
specific conductance (µs/cm)	-d	+#	-a		
lab pH	+c	+#	+#		+b
chloride (mg/*l*)	+#	-#	+b		
ammonia - N (mg/*l*)	-c	+#	-b		+#
inorganic -N (mg/*l*)	-d	-#	-a		+#
filtered total Kjeldahl N	-#	+#	-b		-#
Kjeldahl organic N (mg/*l*)	+#	+#	-b		-#
chlorophyll a	-b	-d			
chlorophyll b	-a	+#			
phytoplankton density					
filtered reactive phosphate	-b	-#	-#		-#
total phosphorous (mg/*l*)	-#	-c	-c		-#
iron (mg/*l*)	-d	+#	-#		-#

Note:

1. a, b, c, d denote significance levels of 10, 5, 1, 0.1 percent, respectively.
2. # denotes result not significant at the 10 percent level.
3. Otherwise, a blank indicates insufficient data.
4. + or - according as the median of X is > or < the median of Y.

Table 23.5.8 presents all of the results of the Wilcoxon tests for all of the variables and stations for which there are sufficient data. Notice that for Stations 501 and 994 the test results for most of the water quality variables are significant. For example, for iron at Station 501 the significance level is $d = 0.1\%$. Hence, one should reject the null hypothesis that the means are the same at the two depths. The negative sign indicates that the mean or median at the shallow depth is less than the mean for samples taken at the deeper depths. At Stations 810 and 1086, the means appear to be the same at the shallow and deep depths for most of the water quality variables.

Kruskal-Wallis Tests

The Kruskal-Wallis test outlined in Appendix A23.3 is a nonparametric test for checking whether or not the means among replicated samples are significantly different from one another. Although the results are not shown here, the one way analysis of variance constitutes a parametric approach for performing the same test but under stricter assumptions.

In Table 23.5.9, the studies for the replicated samples for the total phosphorous data (mg/*l*) at Station 501 are presented. At the top of the table, the number of replicated samples by month and year are given. Below this, the Tukey 5-number summary of all depths is displayed. For a given replicated sample, the range is defined as the largest minus the smallest value. The Tukey 5-number summary for the ranges of all the replicated samples is also given in Table 23.5.9.

Table 23.5.9. Analyses of replicated samples for total phosphorous (mg/*l*) at Station 501, Long Point Bay, Lake Erie.

Number of replicated samples available by month and year

	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	TOTAL
Jan																0
Feb																0
Mar																0
Apr													4	2		6
May													4	2		6
Jun													4	2		6
Jul													6	1		7
Aug													2			2
Sep													6	2		8
Oct													2	3		5
Nov													4	4		8
Dec													2	1		3
TOTAL	0	0	0	0	0	0	0	0	0	0	0	0	34	17	0	51

Tukey five number summary of all the depths(m) used

1.50	1.50	1.50	11.00	11.50

Tukey five number summary of the ranges of the replicates

0.000	0.002	0.003	0.007	0.110

Kruskal-Wallis Nonparametric Test

KW-Statistic	SL
109.31	7.15×10^{-7}

Note that if all the entries in a given replicated sample were the same and this were true for all the replicated samples, all of the entries in the Tukey 5-number summary would be zero.

The results for the Kruskal-Wallis test are given at the bottom of Table 23.5.9. Using [A23.3.3], the test statistic is found to be 109.31 with a SL of 7.15×10^{-7}. Because of the very small SL, the statistic is significant and hence one could argue that the means for two or more replicated samples are significantly different from one another.

The reader should keep in mind that the differences among the means for the replicated samples could be due to causes such as seasonality, trend and depth. Seasonality may be the main reason for the mean differences but more data would be required to test this hypothesis. The fact that the samples are not independent may also influence the results.

The overall results for when the Kruskal-Wallis test is applied to all the time series having sufficient data across all five sites, are shown in Table 23.5.10. From this table, it can be seen that for iron the means among replicated samples are significantly different from one another for iron across all stations for which there are enough data. However, these differences could be due to causes such as seasonality trend or depth.

Table 23.5.10. Kruskal-Wallis tests for comparing means among replicated samples at Long Point Bay, Lake Erie.

	Stations				
	501	810	994	1085	1086
turbidity (FTU)	a		#		#
specific conductance (μs/cm)	#		#		d
lab pH			#		#
chloride (mg/*l*)					
ammonia - N (mg/*l*)	#		a	#	#
inorganic -N (mg/*l*)			#		#
filtered total Kjeldahl N	c		#	a	#
Kjeldahl organic N (mg/*l*)	#		a	#	b
chlorophyll a	#				
chlorophyll b	#				
phytoplankton density	#	#	#		
filtered reactive phosphate			#		#
total phosphorous (mg/*l*)	d		d	d	c
iron (mg/*l*)	d		d	d	d

Note:

1. a, b, c, d denote significance levels of 10, 5, 1, 0.1 percent, respectively.
2. # denotes result not significant at the 10 percent level.
3. Otherwise, a blank indicates insufficient data.

23.6 CONCLUSIONS

Environmental data, such as water quality time series, are often very messy. For example, water quality time series may possess problems which include having missing observations, following nonnormal distributions, possessing outliers, and being short in length. Because nonparametric tests usually have less restrictive assumptions than their parametric counterparts, nonparametric tests are often ideally suited for detecting characteristics such as trends in environmental data (Helsel, 1987). Furthermore, because of the increasing importance of environmental impact assessment studies in modern day society, the import of both nonparametric and parametric tests will continue to expand.

Following a general discussion of statistical testing in Section 23.2, in Section 23.3 a number of useful nonparametric tests are described for detecting trends in data sets. In particular, the seasonal Mann-Kendall test and the correlated seasonal Mann-Kendall test of Section 23.3.2 constitute important intrablock methods for discovering trends in time series for which there may be missing observations. However, the aligned rank technique discussed in the latter part of Section 23.3.2 must be used with an evenly spaced time series.

When dealing with seasonal data measured at one or more sites, the procedures described in Section 23.3.3 can be used for grouping data together when checking for the presence of trends. For instance, if it is suspected that there is an increasing trend during the summer seasons and a decreasing trend at other times of the year, the data can be subdivided into the summer and nonsummer groups. A good way to combine tests of hypotheses across seasons or groups of seasons is to employ Fisher's method given in [23.3.30] in Section 23.3.4. Besides grouping the data, sometimes it is worthwhile to first filter the given time series in order to account for the effect of water quantity upon water quality. One particular approach for filtering or preprocessing data before they are subjected to statistical testing is briefly outlined in Section 23.3.5 while another procedure is presented in Section 24.3.2. Moreover, the Spearman partial rank correlation test of Section 23.3.6 provides a more flexible approach for filtering out undesirable effects when checking for trends over time in a series.

The only nonparametric test that is used to model a step trend due to a known intervention is described in Section 23.3.7. For a given physical variable, this test can be used to confirm the presence of a step trend across seasons at one or more measuring locations. However, it cannot be used to test for the presence of a trend which does not take place as a step change in the mean level after the date of occurrence of a known intervention. Recall from Section 22.4 and Chapter 19 that intervention analysis can be used to model a wide spectrum of trend shapes caused by one or more interventions and also accurately estimate the magnitudes of the trends.

The ACF at lag one is a parametric test which can be used for detecting trends in a data set. On the other hand, Kendall's tau in [23.3.5] or, equivalently, the Mann-Kendall statistic in [23.3.1] or [23.3.7], constitute statistics that can be used in the nonparametric Mann-Kendall test for trend detection. The simulation experiments executed in Section 23.4 demonstrate that the ACF at lag one is more powerful than Kendall's tau for discovering purely stochastic trends while Kendall's tau is more powerful for uncovering purely deterministic trends.

Often water quality and other types of time series are multiple censored. In order to be able to apply nonparametric tests to multiple censored data, one can employ procedures described in Section 23.3.8.

To clearly demonstrate the efficacy of employing nonparametric and also parametric methods in a complex environmental impact assessment study, the effects of industrial development upon water quality in Long Point Bay in Lake Erie, are systematically examined in Section 23.5. The specific statistical methods used in the application for exploratory and confirmatory data analyses are listed in Table 23.5.1. Within Section 23.5.2, the method of application and representative results are given for each of the techniques marked by a cross in Table 23.5.1. Of particular importance is the seasonal Mann-Kendall test that is used to check for the presence of trends in a range of water quality variables at different sites (see Tables 23.5.4 to 23.5.6). Other nonparametric tests utilized in the study include the Wilcoxon signed rank and Kruskal-Wallis tests described in Appendices A23.2 and A23.3, respectively. When deciding upon which tests to employ in a given study, it is informative to refer to tables that characterize statistical methods according to various criteria. Table 23.1.1 summarizes the purpose of all of the nonparametric tests described in Chapter 23.

Beyond applications given in Sections 23.5.2 and 24.3, as well as references already cited in this chapter, there is, of course, other published literature dealing with nonparametric modelling in water resources and environmental engineering. For example, Fox et al. (1990), Potter (1991), and El-Shaarawi and Niculescu (1993) apply nonparametric trend tests to water quantity problems, El-Shaarawi et al. (1983, 1985), Smith et al. (1987), Alexander and Smith (1988), Karlsson et al. (1988), Loftis and Taylor (1989), Lettenmaier et al. (1991), Sanden et al. (1991), Walker (1991), Zetterqvist (1991), and Tsirkunov et al. (1992), employ nonparametric tests for detecting trends in water quality time series, and Hamed and Davenport (1990), Wiseman et al.

(1990), Jordan et al. (1991), and Stanley (1993) apply nonparametric trend tests to estuarine data. In fact, assessment of water quality is of national concern to many countries throughout the world, including the United States of America (see, for instance, Cohen et al., 1988). Consequently, the need for further developing flexible nonparametric trend tests, as well as many other kinds of statistical procedures, will continue to expand.

In the last row of Table 23.5.1, it is noted that regression analysis is used to model some problems related to the water quality study on Lake Erie reported in Section 23.5. Regression analysis is, in fact, a very flexible and general tool that has wide applicability in water resources and environmental engineering. Consequently, in the next chapter various types of regression models are put forward for use as exploratory and confirmatory data analysis tools. Additionally, an overall methodology is presented for systematically carrying out a trend assessment study with messy water quality data measured in a river. When dealing with water quality data from a river, the flow levels must be accounted for and regression analysis provides a superb means for doing this. As is shown in Chapter 24, regression analysis techniques, along with some nonparametric trend tests as well as other statistical methods, play a key role in this methodology.

APPENDIX A23.1

KENDALL RANK CORRELATION TEST

The *Kendall rank correlation test* (Kendall, 1975) is a nonparametric test for checking if two series are independent of one another. The null hypothesis, H_0, is that the two series are independent of each other while the alternative hypothesis, H_1, is they are not independent.

Suppose that the data consist of a bivariate random sample if size n, (x_i, y_i), for $i = 1, 2, \ldots, n$. Two observations are concordant if both members of one bivariate observation are larger than their respective members of the other observation. For example, the two bivariate observations (3.2,9.6) and (4.7,11.2) are concordant. Out of the $\binom{n}{2}$ total possible pairs, let N_c denote the number of concordant pairs of observations. A pair of bivariate observations, such as (5.2,8.6) and (4.3,12.4), is discordant if it is not concordant. Let N_d be the total number of discordant pairs. Under H_0, the test statistic for the Kendall rank correlation test is

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)} \qquad [A23.1.1]$$

If all pairs are concordant, the two series are perfectly correlated and $\tau = 1$. For the case of total discordance, $\tau = -1$. Consequently, τ varies between -1 and +1. Because τ is asymptotically normally distributed and its distribution can be tabulated exactly for small n, one can determine the SL for a computed value of τ. If the calculated τ is greater than or less than 0.05, one can accept or reject, respectively, the null hypothesis. Theoretical results regarding this test are provided by Valz et al. (1994).

Notice that the symbol for tau used in [A23.1.1] is identical to that used in [23.3.5] in Section 23.3.2 for the Mann-Kendall trend test. This is because the test statistic given in [23.3.5] is a special case of the test statistic for the Kendall rank correlation test in [A23.1.1]. To obtain [23.3.5] from [A23.1.1], simply replace (x_i,y_i) by (t,x_t) for which time $t = 1,2, \ldots, n$, and x_t consists of $x_1,x_2, \ldots, x_n$.

APPENDIX A23.2

WILCOXON SIGNED RANK TEST

Wilcoxon Test: In water quality modelling, one may wish to know whether or not measurements taken at two different depths at exactly the same time possess the same median. Moreover, as explained and demonstrated in Sections 8.3 and 15.3, the Wilcoxon signed rank test can be employed for checking whether one time series models forecasts significantly better than another. A test proposed by Wilcoxon (1945) and described in detail by Conover (1980, pp. 280-288) can be used with paired data. Let the data consist of n' pairs of observations $(x_1,y_1),(x_2,y_2), \ldots, (x_{n'},y_{n'})$ generated by their respective bivariate random variables $(X_1,Y_1),(X_2,Y_2), \ldots, (X_{n'},Y_{n'})$. For each of the n' pairs, (X_i,Y_i), the absolute differences can be computed using

$$|D_i| = |Y_i - X_i| \quad i = 1,2, \ldots, n' \qquad [A23.2.1]$$

In the test, pairs for which $X_i = Y_i$ and hence $D_i = 0$ are omitted. Let $n \leq n'$ denote the number of remaining pairs. The n pairs can then be ranked from 1 to n where rank 1 is given to the pair with the smallest $|D_i|$ and rank n is assigned to the pair with the largest $|D_i|$. When there are ties among the absolute differences for a set of paired values, each of the pairs in the set is assigned the average of the ranks that otherwise would have been assigned. As pointed out by Conover (1971, p. 281), the assumptions underlying the Wilcoxon test are each D_i is a continuous random variable, the distribution of each D_i is symmetric, the D_i's are mutually independent, all of the D_i's possess the same median, and the measurement scale of the D_i's is at least interval.

Let d_{50} be the median of the D_i's. For a two-tailed Wilcoxon test, the null hypothesis is

$$H_0{:}d_{50} = 0$$

This implies that the medians of the X_i's and Y_i's are the same. The alternative hypothesis is that the medians of the X_i's and Y_i's are different. This can be written as

$$H_1{:}d_{50} \neq 0$$

Because the distribution of each D_i is assumed to be symmetric, the median is identical to the mean and one is also testing whether or not the means of the X_i's and Y_i's are the same.

The test statistic, T, used to decide if H_0 should be accepted or rejected is defined to be the sum of the ranks assigned to those pairs (X_i,Y_i) where Y_i exceeds X_i. Therefore, for each pair (X_i,Y_i)

$$R_i = \begin{cases} 0 & \text{if } X_i > Y_i \\ \text{rank assigned to } (X_i,Y_i) & \text{if } X_i < Y_i \end{cases} \qquad \text{[A23.2.2]}$$

and the test statistic is written as

$$T = \sum_{i=1}^{n} R_i \qquad \text{[A23.2.3]}$$

Because the exact distribution of T is known, one can easily calculate the SL for T (Conover, 1971, pp. 211-215, p. 383). If, for example, the value of T is either sufficiently large or small enough to cause the significance level to be less than say 5%, one can reject H_0 and thereby assume that the medians of the X_i's and Y_i's are different.

Confidence Interval for the Median: The Wilcoxon signed rank test is employed to check whether or not the median of the X_i's and Y_i's are significantly different from one another. In the test, one actually checks whether or not the median of D_i is significantly different from zero. To obtain an estimate of the magnitude of the unknown median of the D_i's, one can calculate a confidence interval for this median using the method given by Tukey (1949) and also described by Walker and Lev (1953, p. 445) and Conover (1980, pp. 288-290). Moreover, the confidence interval for the median is also a confidence interval for the mean difference

$$E(D_i) = E(Y_i) - E(X_i)$$

if (X_i,Y_i) or D_i, $i = 1,2, \ldots, n$, is a random sample and if the mean difference exists.

To calculate a confidence limit for D_i first one must select a significance level α which means the confidence interval is $1-\alpha$. From the tables for the exact distribution of D_i (Conover, 1980, p. 460-461) one can obtain the $\alpha/2$ quantile denoted by $\omega_{\alpha/2}$. Next, determine the $n(n+1)/2$ possible averages $(D_i + D_j)/2$ for all i and j, including $i = j$. The upper and lower bounds for the $1-\alpha$ confidence interval are given by the $w_{\alpha/2}$th largest of the averages and the $w_{\alpha/2}$th smallest of the averages, respectively. Because of this, one only has to compute the averages near the largest and smallest D_i's and not the entire $n(n+1)/2$ averages.

APPENDIX A23.3

KRUSKAL-WALLIS TEST

The Kruskal-Wallis test constitutes a nonparametric approach for checking whether or not the distributions or means across k samples are the same (Kruskal and Wallis, 1952). In contrast to the normality assumption for the one way analysis of variance (ANOVA), the k population distributions are only assumed to be identical for the Kruskal-Wallis test. However, as is the case for the one way ANOVA, the observations are assumed to be independent of another.

Suppose that there are k samples of sizes n_i, $i = 1,2,\ldots,k$, having a combined sample size n defined by

$$\sum_{i=1}^{k} n_k = n \qquad [A23.3.1]$$

Let x_{ij} stand for the jth value in the ith sample so that $i = 1,2,\ldots,k$, and $j = 1,2,\ldots,n_i$. Denote the ith random sample of size n_i by $x_{i1},x_{i2},\ldots,x_{in_i}$. Rank the n observations from 1 to n where ranks 1 and n are assigned to the smallest and largest observations, respectively. For tied observations, assign each observation the average of the ranks that would be assigned to the observations. Let $R(x_{ij})$ represent the rank assigned to x_{ij}. For the ith sample, the sum of the ranks is given by

$$R_i = \sum_{j=1}^{n_i} R(x_{ij}) \quad i = 1,2,\ldots,k \qquad [A23.3.2]$$

The null hypothesis is that all of the k population distribution functions are identical. This implies that the k means are the same. The alternative hypothesis is that at least one of the populations yields larger observations than at least one of the other populations. Therefore, the k populations do not all have identical means.

The test statistic for the Kruskal-Wallis test is

$$T = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{[R_i - (1/2)n(n+1)]^2}{n_i} \qquad [A23.3.3]$$

where n and R_i are defined in [A23.2.1] and [A23.2.2], respectively. The exact distribution of T is known (Kruskal and Wallis, 1952) and for small samples (say $k = 3$ and $n_i \leq 5$, $i = 1,2,3$) one can obtain the SL for the observed T from tables. For larger samples, T is approximately χ^2 distributed on $k - 1$ degrees of freedom.

Compared to the usual parametric F test used in the one way ANOVA, the Kruskal-Wallis test is very efficient. For example, when the assumptions of the F test are satisfied, the asymptotic relative efficiency of the Kruskal-Wallis test compared to the F test is 0.955. A detailed description of the Kruskal-Wallis test is presented by Conover (1980, pp. 229-237).

The Kruskal-Wallis test can be employed for testing whether or not a time series is seasonal. To apply the test, the data contained within a given season is considered as one separate sample. For example, when dealing with monthly observations, each of the twelve samples would consist of the monthly data from that month across all of the years. A significantly large value of the Kruskal-Wallis statistic would mean that the means and perhaps other distributional parameters vary across the seasons. On the other hand, if the Kruskal-Wallis statistic were not significant, this would mean that the data are not seasonal.

PROBLEMS

23.1 In Section 23.3.2, the correlated seasonal Mann-Kendall statistic is presented as a nonparametric approach for use in trend detection. Describe some research projects that you think would improve this nonparametric test from both theoretical and practical viewpoints when it is used to discover trends in seasonal water quality data.

23.2 Compare the relative advantages and disadvantages of using parametric and non-parametric tests in trend detection and modelling. If you address points that are raised in Chapter 23, give more depth to your explanations than those presented in this chapter. Also, explain some new points of comparison which are not addressed in Chapter 23. Properly reference the sources of your information.

23.3 In Section 23.3.3, procedures are given for grouping seasons for use in trend detection. Compare the relative advantages and disadvantages of the different approaches. Outline a method of grouping which is not given in Section 23.3.3.

23.4 Without referring to the published literature, prove that the Mann-Kendall test statistic S in [23.3.1] is asymptotically normally distributed with the mean and variance given in [23.3.2]. Why is it necessary to know the distribution of S?

23.5 How can one calculate the exact distribution of the Mann-Kendall test statistic, S, in small samples?

23.6 Select a nonseasonal time series which you suspect may contain a trend. Using the Mann-Kendall test, carry out a formal hypothesis test to ascertain if your suspicions are correct. Comment upon your results.

23.7 Outline how the covariance eigenvalue method of Lettenmaier (1988) works for handling correlation among seasons when employing the correlated seasonal Mann-Kendall test. Describe the advantages and drawbacks of this technique when compared to its competitors.

23.8 Choose a seasonal environmental time series that is of interest to you. Employ the seasonal Mann-Kendall test to check for the presence of trends. Be sure to determine how a trend behaves separately within each season and only group seasons together in a meaningful way when carrying out the trend test across seasons. Be certain to emphasize the most interesting results when explaining your findings.

23.9 Using equations, outline the four methods of combining independent tests of hypothesis which are compared by Littell and Folks (1971). Summarize the advantages and drawbacks of each of the four approaches.

23.10 By referring to the research of Hirsch et al. (1982) and Smith et al. (1982), use equations to outline the procedure for determining the flow adjusted concentration (FAC) for a water quality variable. Appraise their approach for calculating the FAC's.

23.11 By employing a seasonal environmental time series, demonstrate how the Spearman partial rank correlation test is used for trend detection when the effects of seasonality are partialled out. Clearly, explain how your calculations are carried out when applying this test and comment upon any interesting results.

23.12 Explain how the Spearman partial rank correlation test can take into account the effects of correlation when checking for the presence of a trend in a time series.

23.13 Design comprehensive simulation experiments to compare the powers of the seasonal Mann-Kendall and Spearman partial rank correlation tests for trend detection.

23.14 Carry out a sensible portion of the simulation experiments designed in the previous problem.

23.15 Define the Kendall partial rank correlation statistic for three variables labelled as X, Y and Z. Explain how this statistic could be utilized in trend tests. Discuss the advantages and disadvantages of using the Kendall partial rank correlation coefficient for discovering trends.

23.16 Define the Pearson partial rank correlation coefficient for three variables denoted as X, Y and Z. Explain various ways in which this statistic could be used for trend tests. Comment upon the advantages and drawbacks of employing the Pearson partial rank correlation coefficient for trend detection.

23.17 In Section 23.3.7 it is noted that Hirsch and Gilroy (1985) employ a filter for determining a filtered sulphate loading series which can then be tested for the presence of a step trend which started at a known intervention date. Using equations, outline how this filter works and explain its advantages and limitations.

23.18 Using equations, explain how the nonseasonal Mann-Kendall test is applied to a multiple censured data set by employing the expected rank vector approach of Hughes and Millard (1988). Using either an actual or hypothetical time series having at best three levels of censoring on the left, demonstrate how the trend test is carried out.

23.19 Execute the instructions of the previous problem for the case of a seasonal time series.

23.20 Define one more deterministic trend model and one more stochastic trend model which could have been used in the simulation studies for trend detection in Section 23.4. Explain the reasons for your choice of models.

23.21 Define a mixed deterministic-stochastic trend model which could be employed in the simulation studies for trend detection presented in Section 23.4. Justify the reasons for your choice of a mixed model and explain why the authors did not use a mixed model in their simulation experiments.

23.22 In Section 23.4, simulation experiments are used to ascertain the ability of the ACF at lag one, r_1, in [23.4.1], and Kendall's tau in [23.3.5] to detect deterministic and stochastic trends. Define one other parametric statistic and another nonparametric statistic which could have been used in the simulation studies. Explain why you selected these statistics and compare them to those used in Section 23.4.

23.23 Prove that the ACF at lag one, r_1, in [23.4.1] is asymptotically NID$(0,\frac{1}{n})$ where n is the sample size.

23.24 A threshold autoregressive model is defined in [23.4.15]. Explain how this model is fitted to a given time series by following the three stages of model construction.

23.25 As noted in Section 23.5.2, both the Kruskal-Wallis test and one way analysis of variance can be used for checking whether or not the means among replicated samples are significantly different from one another. After briefly outlining how each test is designed, compare the merits and drawbacks of the two methods.

23.26 Outline the approach of Montgomery and Reckhow (1984) for carrying out a trend assessment study. Compare their procedure to methodology employed in Section 23.5.

23.27 Summarize the procedure of Hirsch et al. (1991) for selecting methods to employ for detecting and estimating trends in water quality time series.

23.28 Suppose that a government agency gives you a set of environmental time series to examine for the presence of trends. Briefly describe how you would decide upon what to do.

23.29 Obtain some environmental time series which are suspected of possessing trends. List the exploratory and confirmatory data analysis tools which you plan to use and then execute a comprehensive data analysis study.

23.30 Conover and Iman (1981) describe a valuable procedure for linking parametric and nonparametric statistics. Outline how this connection is carried out and summarize the advantages of the approach. Be sure to explain how their procedure can enhance regression analysis, which is the topic of the next chapter.

REFERENCES

BOOKS ON NONPARAMETRIC STATISTICAL METHODS

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs, New Jersey.

Conover, W. J. (1980). *Practical Nonparametric Statistics*. John Wiley, New York, second edition.

Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.

Gibbons, J. D. (1971). *Nonparametric Statistical Inference*. McGraw-Hill, New York.

Gibbons, J. D. (1976). *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston, New York.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Hipel, K. W., editor (1988). *Nonparametric Approaches to Environmental Impact Assessment.* American Water Resources Association (AWRA), AWRA Monograph, Series No. 10, Bethesda, Maryland.

Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. Wiley, New York.

Kendall, M. G. (1975). *Rank Correlation Methods*. Charles Griffin, London, fourth edition.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Oakland, California.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Siegel, S. (1956). *Nonparametric Statistics for Behavioral Sciences*. McGraw-Hill, New York.

CENSORED DATA

Cohn, T. A. (1988). Adjusted maximum likelihood estimation of the moments of lognormal populations from type I censored samples. U.S. Geological Survey Open File Report, pages 88-350.

El-Shaarawi, A. H. (1989). Inferences about the mean from censored water quality data. *Water Resources Research*, 25(4):685-690.

Gilbert, R. O. and Kinnison, R. R. (1981). Statistical methods for estimating the mean and variance from radionuclide data sets containing negative, unreported or less-than values. *Health Physics*, 40:377-390.

Gilliom, R. J. and Helsel, D. R. (1986). Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques. *Water Resources Research*, 22(2):135-146.

Gilliom, R. J., Hirsch, R. M., and Gillroy, E. J. (1984). Effect of censoring trace-level water-quality data on trend-detection capability. *Environmental Science Technology*, 18:530-535.

Gleit, A. (1985). Estimation for small normal data sets with detection limits. *Environmental Science Technology*, 19:1201-1206.

Helsel, D. R. and Cohn, T. A. (1988). Estimation of descriptive statistics for multiple censored water quality data. *Water Resources Research*, 24(12):1997-2004.

Helsel, D. R. and Gilliom, R. J. (1986). Estimation of distributional parameters for censored trace level water quality data, 2. Verification and applications. *Water Resources Research*, 22(2):147-155.

Hirsch, R. M. and Stedinger, J. R. (1987). Plotting positions for historical floods and their precision. *Water Resources Research*, 23(4):715-727.

Hughes, J. and Millard, S. P. (1988). A tau-like test for trend in the presence of multiple censoring points. *Water Resources Bulletin*, 24(3):521-531.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Kushner, E. J. (1976). On determining the statistical parameters for pollution concentration from a truncated data set. *Atmospheric Environment*, 10:975-979.

Lee, E. T. (1980). *Statistical Methods for Survival Data Analysis*. Lifetime Learning, Belmont, California.

Millard, S. P. and Deverel, S. J. (1988). Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. *Water Resources Research*, 24(12):2087-2098.

Miller, R. G. (1981). Survival Analysis. John Wiley, New York.

Owen, W. J. and DeRouen, T. A. (1980). Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants. *Biometrics*, 36:707-719.

Porter, P. S. and Ward, R. C. (1991). Estimating central tendency from uncensored trace level measurements. *Water Resources Bulletin*, 27(4):687-700.

ENCYCLOPEDIAE, HANDBOOK AND DICTIONARY ON STATISTICS

Kendall, M. G. and Buckland, W. R. (1971). *A Dictionary of Statistical Terms*. Longman Group Limited, Thetford, Norfolk, Great Britain, third edition.

Kotz, S. and Johnson, N. L., editors (1988). *Encyclopedia of Statistical Sciences, Volumes 1 to 9*. Wiley, New York.

Kruskal, W. H. and Tanur, J. M. (1978). *International Encyclopedia of Statistics, Volumes 1 and 2*. The Free Press, New York.

Sachs, L. (1984). *Applied Statistics, A Handbook of Techniques*. Springer-Verlag, New York, second edition.

FREQUENCY DOMAIN TREND TESTS

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change*, 21:1-16.

Bloomfield, P. and Nychka, D. (1992). Climate spectra and detecting climate change. *Climatic Change*, 21:275-287.

Bloomfield, P., Oehlert, G., Thompson, M. L. and Zeger, S. (1983). A frequency domain analysis of trends in Dobson total ozone records. *Journal of Geophysical Research*, 88(C13):8512-8522.

NONPARAMETRIC TREND TESTS

Best, D. J. and Gipps, P. G. (1974). The upper tail probabilities of Kendall's tau. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 23(1):98-100.

Crawford, C. G., Hirsch, R. M., and Slack, J. R. (1983). Nonparametric tests for trends in water quality data using the statistical analysis system (SAS). Technical Report U. S. Geological Survey Open-File Report 83-550, U. S. Government.

Dietz, E. J. and Killeen, T. J. (1981). A nonparametric multivariate test for monotone trend with pharmaceutical applications. *Journal of the American Statistical Association*, 76(373):169-174.

El-Shaarawi, A. H. and Niculescu, S. P. (1992). On Kendall's Tau as a test of trend in time series data. *Enviornmetrics*, 3(4):385-411.

El-Shaarawi, A. H. and Niculescu, S. P. (1993). A simple test for detecting non-linear trend. *Environmetrics*, 4(2):233-242.

Farrell, R. (1980). Methods for classifying changes in environmental conditions. Technical Report VRF-EP A7.4-FR80-1, Vector Research Inc., Ann Arbor, Michigan.

Hipel, K. W., McLeod, A. I., and Fosu, P. K. (1986). Empirical power comparisons of some tests for trend. In El-Shaarawi, A. H. and Kwiatkowski, R. E., editors, *Statistical Aspects of Water Quality Monitoring*. Elsevier, Amsterdam.

Hirsch, R. M. (1988). Statistical methods and sampling design for estimating step trends in surface-water quality. *Water Resources Bulletin*, 24(3):493-503.

Hirsch, R. M. and Gilroy, E. J. (1985). Detectability of step trends in the rate of atmospheric deposition of sulphate. *Water Resources Bulletin*, 21(5):773-784.

Hirsch, R. M. and Slack, J. R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, 20(6):727-732.

Hirsch, R. M., Slack, J. R., and Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18(1):107-121.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41:133-145.

Lettenmaier, D. P. (1976). Detection of trends in water quality data from records with dependent observations. *Water Resources Research*, 12(5):1037-1046.

Lettenmaier, D. P. (1988). Multivariate nonparametric tests for trend in water quality. *Water Resources Bulletin*, 24(3):505-512.

Loftis, J. C., Taylor, C. H., and Chapman, P. L. (1991a). Multivariate tests for trend in water quality. *Water Resources Research*, 27(7):1419-1429.

Loftis, J. C., Taylor, C. H., Newell, A. D. and Chapman, P. L. (1991b). Multivariate trend testing of lake water quality. *Water Resources Research*, 27(3):461-473.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13:245-259.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58:216-230.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63:1379-1389.

Smith, R. A., Hirsch, R. M., and Slack, J. R. (1982). A study of trends in total phosphorous measurements at stations in the NASQAN network. Technical Report Water Supply Paper 2190, U. S. Geological Survey, Reston, Virginia.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:88-97.

Taylor, C. H. and Loftis, J. C. (1989). Testing for trend in lake and ground water quality time series. *Water Resources Bulletin*, 25(4):715-726.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, 1, 2, and 3. *Ned. Akad. Wetensch Proc.*, 53:386-392, 521-525, 1397-1412.

Valz, P. (1990). *Developments in Rank Correlation Procedures*. PhD thesis, Department of Statistical and Actuarial Sciences. The University of Western Ontario, London, Ontario, Canada.

Valz, P. D., McLeod, A. I. and Thompson, M. E. (1994). Cumulant generating function and tail probability approximations for Kendall's score with tied rankings. *The Annals of Statistics*.

Van Belle, G. and Hughes, J. P. (1984). Nonparametric tests for trend in water quality. *Water Resources Research*, 20(1):127-136.

Zetterqvist, L. (1988). Asymptotic distribution of Mann's test for trend for m-dependent seasonal observations. *Scandinavian Journal of Statistics*, 15:81-95.

OTHER NONPARAMETRIC TESTS

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124-129.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583-621.

Tukey, J. W. (1949). The simplest signed-rank tests. Mimeographed report No. 17, Statistical Research Group, Princeton University.

Walker, H. M. and Lev, J. (1953). *Statistical Inference*. Holt, New York.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80-83.

SERIAL CORRELATION COEFFICIENT AT LAG ONE

Cox, D. R. (1966). The null distribution of the first serial correlation coefficient. *Biometrika*, 53:623-626.

Dufuor, J. M. and Roy, R. (1985). Some robust exact results on sample autocorrelation and tests of randomness. Technical report, Department of Data Processing and Operations Research, University of Montreal, Montreal, Quebec, Canada.

Kendall, M. G., Stuart, A., and Ord, J. K. (1983). *The Advanced Theory of Statistics, Volume 3*. Griffin, London.

Knoke, J. D. (1975). Testing for randomness against autocorrelated alternatives: The parametric case. *Biometrika*, 62:571-575.

Knoke, J. D. (1977). Testing for randomness against autocorrelation: Alternative tests. *Biometrika*, 64:523-529.

Noether, G. E. (1950). Asymptotic properties of the Wald-Wolfowicz test of randomness. *Annals of Mathematical Statistics*, 21:231-246.

Wald, A. and Wolfowicz, J. (1943). An exact test for randomness in the nonparametric case based on serial correlation. *Annals of Mathematical Statistics*, 14:378-388.

SIMULATION MODELS

Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society, Series B*, 21:239-244.

Cleary, J. A. and Levenbach, H. (1982). *The Professional Forecaster*. Lifetime Learning Publications, Belmont, California.

Cochran, W. G. (1977). *Sampling Techniques*. Third edition, Wiley, New York.

Tong, H. (1977). Discussion of paper by A. J. Lawrence and N. T. Kottegoda. *Journal of the Royal Statistical Society, Series A*, 140:34-35.

Tong, H. (1978). On a threshold model. In Chen, C. H., editor, *Pattern Recognition and Signal Processing*. Sijthoff and Noordhoff, The Netherlands.

Tong, H. (1983). Threshold models in non-linear time series analysis. *Lecture Notes in Statistics*, No. 21, Springer-Verlag, New York.

Tong, H. and Lim, S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42:245-292.

Tong, H., Thanoon, B., and Gudmundsson, G. (1985). Threshold time series modelling of two Icelandic riverflow systems. *Water Resources Bulletin*, 21(4):651-661.

STATISTICS

Bahadur, R. R. (1967). An optimal property of the likelihood ratio statistic. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, I:13-26.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburg, England.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. 3rd edition, Oliver and Boyd, Edinburg.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley, New York, second edition.

Hipel, K. W. (Editor) (1985). *Time Series Analysis in Water Resources*. American Water Resources Association, Bethesda, Maryland.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method for combining independent tests. *Journal of the American Statistical Association*, 66:802-806.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, Series A*, 20:175-240, 263-294.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*, 231:289-337.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31:9-12.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677-680.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

Zar, J. H. (1974). *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.

TREND ASSESSMENT METHODOLOGIES

Berryman, D., Bobee, B., Cluis, D., and Haemmerli, J. (1988). Nonparametric tests for trend detection in water quality time series. *Water Resources Bulletin*, 24(3):545-556.

Harcum, J. B., Loftis, J. C. and Ward, R. C. (1992). Selecting trend tests for water quality series with serial correlation and missing values. *Water Resources Bulletin*, 28(3):469-478.

Hipel, K. W., McLeod, A. I., and Weiler, R. R. (1988). Data analysis of water quality time series in Lake Erie. *Water Resources Bulletin*, 24(3):533-544.

Hipel, K. W. and McLeod, A. I. (1989). Intervention analysis in environmental engineering. *Environmental Monitoring and Assessment*, 12:185-201.

Hirsch, R. M., Alexander, R. B., and Smith, R. A. (1991). Selection of methods for the detection and estimation of trends in water quality. *Water Resources Research*, 27(5):803-813.

McLeod, A. I., Hipel, K. W., and Bodo, B. A. (1991). Trend analysis methodology for water quality time series. *Environmetrics*, 2(2):169-200.

Montgomery, R. H. and Reckhow, K. H. (1984). Techniques for detecting trends in lake water quality. *Water Resources Bulletin*, 20(1):43-52.

WATER QUALITY AND QUANTITY

Water quality applications are also given in published literature listed under other categories for the references in this chapter.

Alexander, R. B. and Smith, R. A. (1988). Trends in lead concentrations in major U. S. rivers and their relationship to historical changes in gasoline-lead consumption. *Water Resources Bulletin*, 24(3):557-569.

Bodo, B. and Unny, T. E. (1983). Sampling strategies for mass-discharge estimation. *Journal of Environmental Engineering, American Society of Civil Engineers*, 109(4):812-829.

Borman, F. H., Likens, G. E., Siccama, T. G., Pierce, R. S., and Eaton, J. S. (1974). The export of nutrients and recovery of stable conditions following deforestation at Hubbard Brook. *Ecological Monographs*, 44:255-277.

Cohen, P., Alley, W. M., and Wilber, W. G. (1988). National water-quality assessment: Future directions of the U. S. Geological Survey. *Water Resources Bulletin*, 24(6):1147-1151.

El-Shaarawi, A. H., Esterby, S. R., and Kuntz, K. W. (1983). A statistical evaluation of trends in the water quality of the Niagara River. *Journal of Great Lakes Research*, 9(2):234-240.

El-Shaarawi, A. H., Esterby, S. R., Warry, N. D., and Kuntz, K. W. (1985). Evidence of contaminant loading to Lake Ontario from the Niagara River. *Canadian Journal of Fisheries and Aquatic Sciences*, 42:1278-1289.

Fox, J. P., Mongan, T. R., and Miller, W. J. (1990). Trends in freshwater inflow to San Francisco Bay from the Sacramento-San Joaquin Delta. *Water Resources Bulletin*, 26(1):101-116.

Harned, D. A., Daniel III, C. C., and Crawford, J. K. (1981). Methods of compensation as an aid to the evaluation of water quality trends. *Water Resources Research*, 17(5):1389-1400.

Harned, D. A. and Davenport, M. S. (1990). Water-quality trends and basin activities and characteristics for the Albemarle-Pamlico estuarine system, North Carolina and Virginia. *U. S. Geological Survey*, Open File Report, 90-398.

Helsel, D. R. (1987). Advantages of nonparametric procedures for analysis of water quality data. *Journal of Hydrological Sciences*, 32(2):179-190.

Hobbie, J. E. and Likens, G. E. (1973). Output of phosphorous, dissolved organic carbon and fine particulate carbon from Hubbard Brook watersheds. *Limnology Oceanography*, 18(5):734-742.

Johnson, N. M., Likens, G. E., Bormann, F. H., Fisher, D. W., and Pierce, R. S. (1969). A working model for the variation in stream water chemistry at the Hubbard Brook Experimental Forest, New Hampshire. *Water Resources Research*, 5(6):1353-1363.

Jordan, T. E. and Correll, D. L., Miklar, J. and Weller, D. E. (1991). Long-term trends in estuarine nutrients and chlorophyll, and short-term effects of variation in water-shed discharge. *Mar. Ecol. Prog. Ser.*, 75:121-132.

Karlsson, G., Grimvall, A. and Lowgren, M. (1988). Riverbasin perspective on long-term changes in the transport of nitrogen and phosphorous. *Water Research*, 22:139-149.

Langbein, W. B. and Dawdy, D. R. (1964). Occurrence of dissolved solids in surface waters in the United States. Professional Paper 501-D, D115-D117, U. S. Geological Survey.

Lettenmaier, D. P., Hooper, E. R., Wagoner, C., and Faris, K. B. (1991). Trends in stream quality in the continental United States, 1978-1987. *Water Resources Research*, 27(3):327-339.

Loftis, J. C. and Taylor, C. H. (1989). Detecting acid precipitation impacts on lake water quality. *Environmental Management*, 13:529-539.

Potter, K. W. (1991). Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment. *Water Resources Research*, 27(5):845-855.

Reckhow, K. H. (1978). Quantitative techniques for the assessment of lake quality. Technical report, Department of Resource Development, Michigan State University.

Sanden, P., Rahm, L., and Wulff, F. (1991). Non-parametric trend test of Baltic Sea data. *Environmetrics*, 2(3):263-278.

Smith, R. A., Alexander, R. B., and Wolman, M. G. (1987). Water-quality trends in the nation's rivers. *Science*, 235:1607-1615.

Stanley, D. W. (1993). Long-term trends in Pamlico River estuary nutrients, Chlorophpyll, dissolved oxygen, and watershed nutrient production. *Water Resources Research*, 29(8):2651-2662.

Tsirkunov, V. V., Nikanorov, A. M., Laznik, M. M. and Zhu, D. (1992). Analysis of long-term and seasonal river water quality changes in Latvia. *Water Research*, 26(5):1203-1216.

Waite, T. D. (1984). *Principles of Water Quality*. Academic Press, Orlando, Florida.

Walker, W. W. (1991). Water quality trends at inflows to Everglades National Park. *Water Resources Bulletin*, 27(1):59-72.

Wiseman, W. J., Jr., Swenson, E. M. and Power, J. (1990). Salinity trends in Louisiana estuaries. *Estuaries*, 13:265-271.

Zetterqvist, L. (1991). Statistical estimation and interpretation of trends in water quality time series. *Water Resources Research*, 27(7):1637-1648.

CHAPTER 24

REGRESSION ANALYSIS AND TREND ASSESSMENT

24.1 INTRODUCTION

Suppose that one would like to analyze trends in water quality time series measured in rivers. The underlying conceptual model used in *trend analysis* of a water quality variable Y can be written as

$$Y_t^{(\lambda)} = f(X_t) + S_t + C_t + \varepsilon_t \qquad [24.1.1]$$

where Y_t is the water quality observation at time t that may be transformed using the Box-Cox transformation in [3.4.30] to produce the transformed value given as $Y_t^{(\lambda)}$, $f(X_t)$ is a function of a covariate series X_t, X_t is a covariate series at time t such as riverflow or temperature measured at time t, S_t is the seasonal component at time t, C_t is the trend in Y_t, and ε_t is the noise component at time t. In trend analysis, one wishes to appropriately account for X_t, S_t and ε_t so that C_t can be easily detected and accurately quantified, even when the trend effects are small. Trends over time can be increasing, decreasing or non-existent. Furthermore, trends over time can follow linear or nonlinear geometrical patterns. For many water quality series measured in rivers, the covariate series used is flow. However, in other situations, different covariate series can be used. For example, temperature may be better to use as X_t when the Y_t series is dissolved oxygen or total nitrates. Also, when water quality measurement sites are far removed from flow gauging sites, covariates other than flow may have to be used for X_t. Finally, the idea of decomposing a series into its basic components as in [24.1.1] is a well established procedure and is, for example, inherent in the basic designs of the intervention models in [19.5.8] and [22.4.5] as well as the seasonal adjustment procedure of Section 22.2.

The objectives of this chapter are twofold. The first goal is to explain some ways in which *regression analysis* can be used to model various components of the general model in [24.1.1]. The second objective is to develop a general *trend analysis methodology* based on [24.1.1] for analyzing trends in water quality time series measured in rivers. As is explained in Section 24.3.2, regression analysis as well as nonparametric tests play a key role in this methodology. To demonstrate the efficacy of the methodology, it is applied to representative water quality time series measured in rivers in Southern Ontario, Canada.

As pointed out in the preface to Part X as well as Section 23.1, water quality and other kinds of environmental data are often quite *messy*. The time series may, for example, be highly skewed, have many missing observations, possess multiple censoring levels, and contain seasonal trends. When examining environmental data for the presence of trends and other statistical properties, a systems design approach to data analysis should be followed. More specifically, one may wish to adhere to the two *data analysis* stages consisting of exploratory data analysis and confirmatory data analysis (Tukey, 1977). This comprehensive approach to data analysis is described in Sections 1.2.4 and 22.1. Graphical procedures that can be used as *exploratory data analysis* tools for visually discovering the main statistical properties of a data set are presented in Section 22.3 as well as Section 24.2.2 in this chapter. *Confirmatory data analysis* techniques for carrying out hypothesis testing and obtaining rigorous statistical statements about certain statistical properties such as trends are presented throughout the book. These confirmatory tools consist of both parametric models and nonparametric tests. *Parametric models,* such as the intervention models of Chapters 19 and Section 22.4, can be used with data that are not too messy. For example, if there are a few missing observations in a series, the intervention model can be used to estimate these missing values as well as estimate the magnitude of a trend. When the data are very messy, one may have to use *nonparametric methods.* A variety of nonparametric tests for trend detection are given in Section 23.3.

Regression analysis models have a very flexible design and can be used with data that are not evenly spaced over time. In the next section, certain kinds of regression models are described for use as exploratory and confirmatory data analysis tools. A particularly informative regression analysis approach for employment as an exploratory tool for tracing trends is the robust locally weighted regression analysis smooth of Cleveland (1979). This technique is described in detail in Section 24.2.2, while applications of the method are given in that section as well as in 24.3.2.

Table 1.6.4 outlines the three trend analysis methodologies presented in the book. Subsequent to carrying out exploratory data analysis studies and filling in missing observations using the seasonal adjustment method of Section 22.2, intervention analysis is employed in Section 22.4 to describe the impacts of cutting down a forest upon the mean levels of riverflows and water quality. In Section 23.5, water quality applications are used for explaining how exploratory and confirmatory data analysis tools can be employed for studying trends in water quality variables measured in a lake. The purpose of Section 24.3 in this chapter is to present a *general trend analysis methodology for use with water quality time series measured in rivers.* As explained in that section and outlined in Table 24.3.1, the methodology consists of graphical trend studies and trend tests. Different kinds of graphs for observing trends are presented in Sections 22.3 and 24.2.2, while nonparametric trend tests are given in Section 23.3. In Section 24.3, procedures are also described for accounting for the effects of flow or another appropriate covariate upon a given water quality series and eliminating any trend in the flow before its effect upon the water quality series is removed. The *Spearman partial rank correlation test* of Section 23.3.6 provides a powerful test for trend detection in a water quality variable over time when the effects of seasonality or some other factors are partialled out. Water quality data measured in rivers are utilized for illustrating how the trend analysis methodology is applied in practice.

24.2 REGRESSION ANALYSIS

24.2.1 Introduction

Regression analysis constitutes a flexible and highly developed parametric modelling approach which has been applied to virtually every field in which data are measured. A host of books on regression analysis are available including valuable contributions by Mosteller and Tukey (1977), Draper and Smith (1981), Atkinson (1985), and Chambers and Hastie (1992).

Regression models can be designed for modelling many situations, including the type of model in [24.1.1]. In addition to major developments in linear regression models, good progress has been made in nonlinear regression (see, for example, Gallant (1987) and Bates and Watts (1988)).

Consider the case of a linear regression model. When fitting a regression model to a data set, it is recommended to follow the identification, estimation and diagnostic check stages of *model development*, as is also done for all of the time series models presented in this book. A wide variety of well developed procedures are available for constructing linear regression analysis models. Usually, the noise term in a regression model is assumed to be normally independently distributed (NID). If the residuals of a fitted model are not normal, one may be able to rectify the situation by transforming the data using the Box-Cox transformation in [3.4.30]. When the residuals are not independent and hence are correlated, one will have to design a more complicated regression model to overcome this problem or, perhaps, use a completely different type of model such as some kind of stochastic model.

As is also the case for the transfer function noise (TFN) and intervention models of Parts VII and VIII, respectively, a regression model can handle multiple input series. However, recall that in TFN and intervention models, the noise term is correlated and modelled using an ARMA model. In a regression model, the noise term is assumed to be white. Another advantage of a TFN model over a regression model is that the transfer function in a TFN model has an operator in both the numerator and denominator as in [17.2.1] and [17.5.3]. This allows one to handle a wide range of ways in which an input series can effect the output or the response variable, as is discussed at the end of Section 17.2.4 and in Section 19.2.2. Furthermore, this can be done using very few model parameters. In fact, one can think of a TFN model as being a regression model having an autocorrelated noise term rather than white noise. Nonetheless, an advantage of regression analysis over TFN models is that it can be used with data that are not evenly spaced and hence possess missing values.

What is meant by *missing* observations must be explained in more detail, especially for the case of water quality time series (McLeod et al. (1991)). First, consider what missing means with respect to parametric techniques. For many parametric methods, such as the wide variety of time series models given in this book, it is usually assumed that observations are available at equally spaced time intervals. For example, when fitting a periodic autoregressive model in [14.2.1] to average monthly riverflow series, all of the monthly observations across the years must be available. If there is at least one missing observation, the data are no longer evenly spaced due to this missing value. Before fitting the time series model, one must obtain estimates for the missing value or values. Furthermore, for the case of riverflows usually each monthly observation is calculated as a monthly average of average daily flows. Each daily average may be based upon a continuous record taken for that day. Hence, the monthly flow data are often calculated from continuous analogue records.

In contrast to riverflow records, water quality observations usually have quite different meanings in sampling theory. More specifically, most water quality records could be classed as irregular series of *quasi-instantaneous measurements*. This is because each water quality sample takes about 10 to 15 seconds to collect. With such samples, the term *missing data* could refer to a number of situations including:

1. All the unsampled 10 to 15 second periods of the record. This total number is, of course, very large and can be thought of as infinity for practical purposes.
2. Uncollected or lost samples with respect to a specific monitoring objective. For instance, an objective may be to collect one sample per month and if at least one monthly observation is missing the data are irregularly spaced due to this missing value. Another objective may be to have most of the water quality samples taken during times of high flows to produce flow biased data. If no samples are taken for at least one high flow event, then there are missing data.
3. Data missing in the sense of some analytical framework. For example, suppose that the framework chosen is the monthly level. If there is at least one observation per month, then one can say that there are no missing values. However, at a daily level there may be many missing observations.

Another stated characteristic of messy environmental data is that the observations may be significantly affected by *external interventions*. These interventions may be man-induced or natural. An example of a beneficial man-induced intervention is the introduction of tertiary treatment at city sewage plants located in a river basin. This beneficial intervention should cause a step decrease in the phosphorus levels in the river, as shown in Figures 1.1.1 as well as 19.1.1 and explained in Section 19.4.5. On the other hand, uncontrolled industrial development with few environmental controls would cause detrimental impacts upon certain water quality variables in a river. One can cite many other examples of environmental policy and related land use changes which can adversely or beneficially affect water quality. An illustration of a natural intervention is the effect of a forest fire in a river basin upon water quality variables. For example, the resulting lack of forest cover may cause more sediments to be carried and deposited by rivers. However, as a new forest grows back over the years, the flows and water quality variables may slowly revert to their former states, which is the situation for the intervention analysis application presented in Section 19.5.4. In trend analysis one wishes to detect and analyze trends caused by man-made or natural interventions.

The intervention model of Part VIII and Section 22.4 constitutes a flexible type of model that can be used to estimate trends. Regression models can also be designed for estimating the magnitude of trends in a series. One approach is to model all of the components in [24.1.1] employing a regression model. Alternatively, one could describe some of the components in [24.1.1] using regression analysis and the remaining parts utilizing other statistical methods. For instance, a nonparametric trend test (see Section 23.3) could be applied to the residuals of a regression analysis to check for the presence of trends after the covariate and seasonality points have been suitably accounted for using regression analysis (see Section 23.3.5). Brown et al. (1975) employ cumulative sum statistics for tests of change in a regression model structure over time. Within the water resources and environmental engineering literature applications of regression analysis in trend assessment include contributions by Alley (1988), Cunningham and Morton (1983), El-Shaarawi et al. (1983), Loftis et al. (1991), McLeod et al. (1991), Smith and Rose (1991), Reinsel and Tiao (1987), Stoddard (1991), and Whitlatch and Martin (1988).

Esterby and El-Shaarawi (1981a,b) devise a procedure for estimating the *point of change and degree in polynomial regression*, while El-Shaarawi and Esterby (1982) extend the approach for use with a regression model having an autoregressive error process of order one. In environmental engineering problems, the time at which an intervention takes place due to additional pollution loads and other reasons is often unknown. The approach of Esterby and El-Shaarawi can estimate the time of the intervention (called the point of change) as well as the order of the polynomial regression and associated model parameters both before and after the intervention. Moreover, Esterby (1985) provides a flexible computer program which allows practitioners to implement their versatile regression analysis technique. As an exploratory data analysis tool, the method of Esterby and El-Shaarawi (1981a,b) and El-Shaarawi and Esterby (1982) is useful for detecting the presence and start of the effects of an intervention. Because estimates of the model parameters both before and after the start of the intervention are obtained, their technique can also be considered as a confirmatory data analysis tool. Finally, El-Shaarawi and Delorme (1982) present statistics for detecting a change in a sequence of ordered binomial random variables.

As pointed out earlier in Section 19.2.3, MacNeill (1985) also presents a flexible technique for detecting and modelling the effects of unknown interventions. In particular, he develops a procedure called *adaptive forecasting and estimation using change-detection*. To activate this adaptive procedure, a successively updated change-detection statistic is proposed. The larger the change in the parameters in a regression, exponential smoothing, ARMA or other type of model, the larger is the expected value of the change-detection statistic. Additional research related to MacNeill's change-detection statistic is referred to in Section 19.2.3 under other trend detection techniques. In addition to trend assessment, regression analysis has been extensively utilized for addressing a wide variety of problems arising in water resources and environmental engineering. For example, interesting applications of regression analysis to natural phenomena are provided by Beauchamp et al. (1989), Cleaveland and Durick (1992), Cohn et al. (1992), Duffield et al. (1992), Gunn (1991), Helsel and Hirsch (1992), Keppeler and Ziemer (1990), Kite and Adamowski (1973), Lyman (1992), Millard et al. (1985), Porter and Ward (1991), Potter (1991), See et al. (1992), Simpson et al. (1993), Tasker (1986), Wong (1963), and Wright et al. (1990). Moreover, fuzzy regression analysis (Bardossy, 1990; Kacrzyk and Federizzi, 1992; Tanaka et al., 1982), has been employed in environmental applications (Bardossy et al., 1990, 1992).

Regression models can be used as both exploratory and confirmatory data analysis tools. Section 24.2.2 presents a flexible regression model for visualizing trends in a series at the exploratory data analysis stage. In Section 24.2.3, an example of designing a regression model as a confirmatory tool in an environmental study is presented. Finally, for discussions about the potential pitfalls that one should be aware of when applying regression analysis as well as other kinds of statistical techniques, the reader may wish to refer to the references cited in problem 24.4 at the end of the chapter.

24.2.2 Robust Locally Weighted Regression Smooth

Overview

Suppose that two variables that can be samples are denoted by X and Y. The measurements for these variables are given by (x_i,y_i), $i = 1,2, \ldots, n$. In a scatterplot, the values for the X and Y variables can be plotted as the abscissae and ordinates, respectively. To gain insight into the relationship between X and Y, it is informative to plot some type of smoothed curve through the scatterplot. In the final part of this section, it is explained how smoothed curves can be obtained for a graph of a single time series and also a scatterplot of a time series where values at time $t-k$ are plotted against those at time t. However, for convenience and generality of presentation, developing a smoothed curve for (x_i,y_i) is now discussed.

A flexible type of smoothing procedure which works well in practice is the *robust locally weighted regression smooth (RLWRS)* developed by Cleveland (1979). Cleveland (1979, 1985), Chambers et al. (1983) and others refer to the general smoothing procedure as LOESS or LOWESS for locally weighted least square regression (when this procedure is iterated robustness is taken into account). Whatever the case, in this chapter the acronym RLWRS is employed. The RLWRS is a member of a set of regression procedures that are commonly referred to as *nonparametric regression* (Stone, 1977). In practice, the RLWRS has been applied to a rich range of problems across many fields. For example, Bodo (1989) and McLeod et al. (1991) have utilized RLWRS for trend assessment of water quality time series, and the RLWRS is also applied to water quality time series in Sections 24.2.2 and 24.3.2 in this chapter. Moreover, Cleveland et al. (1990) have developed a seasonal-trend decomposition procedure based upon the RLWRS.

In essence, the RLWRS is a method for smoothing a scatterplot of (x_i,y_i), $i = 1,2, \ldots, n$, in which the fitted value at x_k is the value of a polynomial fitted to the data using weighted least squares. The weight for (x_i,y_i) is large if x_i is close to x_k and is small if this is not the case. To display graphically the RLWRS on the scatterplot of (x_i,y_i), one plots $(x_i,\hat{y}_i)$ on the same graph as the scatterplot of (x_i,y_i), where $(x_i,\hat{y}_i)$ is called the smoothed point at x_i and $\hat{y}_i$ is called the fitted value at x_i. To form the RLWRS, one simply joins successive smoothed points $(x_i,\hat{y}_i)$ by straight lines. Because a robust fitting procedure is used to obtain the RLWRS, the smoothed points are not distorted by extreme values or other kinds of deviant points.

General Procedure

As explained by Cleveland (1979), the general idea behind his smoothing procedure is as follows. Let W be a weight function which has the following properties:

1. $W(x) > 0$ for $|x| < 1$.
2. $W(-x) = W(x)$.
3. $W(x)$ is a nonincreasing function for $x \geq 0$.
4. $W(x)=0$ for $|x|\geq 1$.

If one lets $0 < f \leq 1$ and r be $(f \cdot n)$ rounded to the nearest integer, the outline of the procedure is as given below. For each x_i, weights, $w_k(x_i)$, are defined for all x_k, $k = 1,2, \ldots, n$, by employing the weight function W. To accomplish this, center W at x_i and scale W so that the point at which W first becomes zero is the rth nearest neighbour of x_i. To obtain the initial fitted value, $\hat{y}_i$, at each x_i a dth degree polynomial is fitted to the data using weighted least squares with weights $w_k(x_i)$. This procedure is called *locally weighted regression*. Based upon the size of the residual $y_i - \hat{y}_i$, a different weight, δ_i, is defined for each (x_i,y_i). In general, large residuals cause small weights while small residuals result in large weights. Because large residuals produce small weights, the effects of extremes tend to be toned down or smoothed, thereby making the procedure *robust*. After replacing $w_k(x_i)$ by $\delta_i w_k(x_i)$, new fitted values are computed using locally weighted regression. The determination of new weights and fitted values is repeated as often as required. All of the foregoing steps taken together are referred to as *robust locally weighted regression*.

In the smoothing procedure, points in the neighbourhood of (x_i,y_i) are used to calculate $\hat{y}_i$. Because the weights $w_k(x_i)$ decrease as the distance of x_k from x_i increases, points whose abscissae are closer to x_i, have a larger effect upon the calculation of $\hat{y}_i$ while further points play a lesser role. By increasing f, the neighbourhood of points affecting $\hat{y}_i$ becomes larger. Therefore, larger values of f tend to cause smoother curves.

In the RLWRS procedure, *local regression* means that regression at a given point is carried out for a subset of nearest neighbours such that the observations closer to the specified point are given larger weights. By taking the size of the residuals into account for obtaining revised weights, robustness is brought into the procedure. Finally, the robust locally weighted regression analysis is carried out for each observation.

Specific Procedure

The procedure presented by Cleveland (1979) for determining the RLWRS is as follows:

1. First the weight function, W, must be specified. Let the distance from x_i to the rth nearest neighbour of x_i be denoted by h_i for each i. Hence, h_i is the rth smallest number among $|x_i - x_j|$, for $j = 1,2,\ldots,n$. For $k = 1,2,\ldots,n$, let

$$w_k(x_i) = W((x_k - x_i)/h_i) \qquad [24.2.1]$$

A possible form for the weight function, is the tricube given by

$$\begin{aligned} W(x) &= (1 - |x|^3)^3 \quad \text{for } |x| < 1 \\ &= 0 \qquad\qquad\quad \text{for } |x| > 1 \end{aligned} \qquad [24.2.2]$$

2. The second step describes how locally weighted regression is carried out. For each i, determine the estimates, $\hat{\beta}_j(x_i)$, $j = 0,1,\ldots,d$, of the parameters in a polynomial regression of degree d of y_k on x_k. This is fitted using weighted least squares having weight $w_k(x_i)$ for (x_k,y_k). Therefore, the $\hat{\beta}_j(x_i)$ are the values of β_j which minimize

$$\sum_{k=1}^{n} w_k(x_i)(y_k - \beta_0 - \beta_1 x_k - \beta_2 x_k^2 - \cdots - \beta_d x_k^d)^2 \qquad [24.2.3]$$

When using locally weighted regression of degree d, the smoothed point at x_i is $(x_i,\hat{y}_i)$ for which $\hat{y}_i$ is the fitted value of the regression at x_i. Hence,

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(x_i) x_i^j = \sum_{k=1}^{n} r_k(x_i) y_k \qquad [24.2.4]$$

where $r_k(x_i)$ does not depend on y_j, $j = 1,2,\ldots,n$. Cleveland (1979) uses the notation $r_k(x_i)$ to reinforce the fact that the $r_k(x_i)$ are the coefficients for the y_k coming from the regression.

3. Let the bisquare weight function be given by

$$B(x) = (1 - x^2)^2 \quad \text{for } |x| < 1$$
$$= 0, \qquad \text{for } |x| \geq 1 \qquad [24.2.5]$$

Let the residuals for the current fitted values be $e_i = y_i - \hat{y}_i$. The robustness weights are defined by

$$\delta_k = B(e_k/6s) \qquad [24.2.6]$$

where s is the median of the $|e_i|$. As pointed out by Cleveland (1979), other types of weight functions could be used in place of $B(x)$.

4. This step is used to calculate an iteration of robust locally weighted regression. For each i, determine new $\hat{y}_i$ by fitting a dth degree polynomial using weighted least squares having the weight $\delta_k w_k(x_i)$ at (x_k, y_k).
5. Iteratively execute steps 3 and 4 for a total of t' times. The final $\hat{y}_i$ constitute the fitted values for the robust locally weight regression and the $(x_i, \hat{y}_i)$ $i = 1, 2, \ldots, n$, form the RLWRS.

Selecting Variables

In order to employ the above procedure, one must specify f, d, t' and W. Cleveland (1979) provides guidelines for doing this. First consider the variable f, where $0 < f \leq 1$, which controls the amount or level of smoothness. As noted earlier, an increase in f causes an increase in the smoothness of the RLWRS. The objective is to select a value of f which is as large as possible to minimize the variability in the smoothed points but without hiding the fundamental pattern or relationship in the data. When it is not certain which value of f to select, setting $f = 0.5$ often produces reasonable results. In practice, one can experiment with two or three values of f and select the one which produces the most informative smooth. Bodo (1989) provides suggestions for selecting f for monthly and lower frequency monitoring data.

Instead of qualitatively selecting one or more values of f, one can estimate f. Based upon the research of Allen (1974), Cleveland (1979) suggests an approach for automatically determining f using a computerized algorithm. The approach begins with the locally weighted regression in step 2. Leaving y_i out of the calculation, for a specified value of f let $\hat{y}_i(f)$ be the fitted value of y_i. A starting value of f_0 for f is chosen by minimizing

$$\sum_{k=1}^{n} (y_k - \hat{y}_k(f))^2 \qquad [24.2.7]$$

Next, using $f = f_0$ the robustness weights in [24.2.6] in step 3 can be determined. Omitting y_i from the calculation and using the robustness weights, let $\hat{y}_i(f)$ be the fitted value at x_i for a given value of f. The next value of f is determined by minimizing

$$\sum_{k=1}^{n} \delta_k (y_k - \hat{y}_k(f))^2 \qquad [24.2.8]$$

Using the latest estimated value of f, the last step can be repeated as many times as are necessary in order to converge to a suitably accurate estimate for f. Depending upon the problem at hand, this procedure for estimating f may require substantial computational time.

The parameter d is the order of the polynomial that is locally fitted to each point. When $d = 1$, a linear polynomial is specified. This usually results in a good smoothed curve that does not require high computational effort and, therefore, a linear polynomial is commonly used.

The parameter t' stands for the number of iterations of the robust fitting procedure. Based upon experimentation, Cleveland (1979) recommends using $t' = 2$. However, the authors of this book have found that $t' = 1$ is sufficient for most applications.

In the description of the general procedure for RLWRS, four required characteristics of the weight function, $W(x)$, are given. It is also desirable that the weight function smoothly decreases to zero as x goes from 0 to 1. The tricube function in [24.2.2] possesses all of the above stipulated properties. Of course, other appropriate weight functions possessing the above attributes could also be entertained.

Applications

The RLWRS can clearly depict meaningful relationships for the situations described below:

1. *Scatter Plot of X against Y* - The measurements of variables X and Y are given by (x_i,y_i), $i = 1,2, \ldots, n$. In a scatter plot, the values for the X and Y variables are plotted as the abscissae and ordinates, respectively. A RLWRS through the scatter plot can visually display the underlying relationship between X and Y.
2. *Time Series Plot* - In a graph of a time series, one plots the values of the time series x_t at each time t against time $t = 1,2, \ldots, n$. Consequently, the point (x_i,y_i) used in the procedure described above is simply replaced by (t,x_t).
3. *Scatter Plot of a Single Time Series* - In a scatter plot of a variable X, one plots x_{t-k} against x_t in order to see how observations separated by k time lags are related. In the procedure for determining the RLWRS, simply substitute (x_t,x_{t-k}) for (x_i,y_i) in order to obtain the smooth.

Applications are now given to illustrate how RLWRS's can be useful in a time series plot and a scatter plot. The data used in the graphs are water quality data from the Ontario Ministry of the Environment for the Saugeen River at Burgoyne, Ontario, Canada.

As explained in Section 22.3.2, one of the simplest and most informative exploratory data analysis tools is to plot the data against time. Characteristics of the data which are often easily discovered from a perusal of a graph include the detection of extreme values, trends due to known or unknown interventions, dependencies between observations, seasonality, need for a data transformation, nonstationarity and long term cycles.

A time series plot is especially useful for visually detecting the presence or absence of a trend. Figure 24.2.1, for example, shows a graph of logarithmic total nitrates for the Saugeen River against time, where each observation is marked using a cross. The fact that $\lambda = 0$ is written above the graph means that the natural logarithmic transformation from [3.4.30] is invoked. The two horizontal lines plotted in the graph delineate the 95% confidence interval (CI) limits if the series is assumed to be normally independently distributed (NID). The observations that lie far outside the 95% CI in Figure 24.2.1 can be considered as outliers under the assumption that the data are NID.

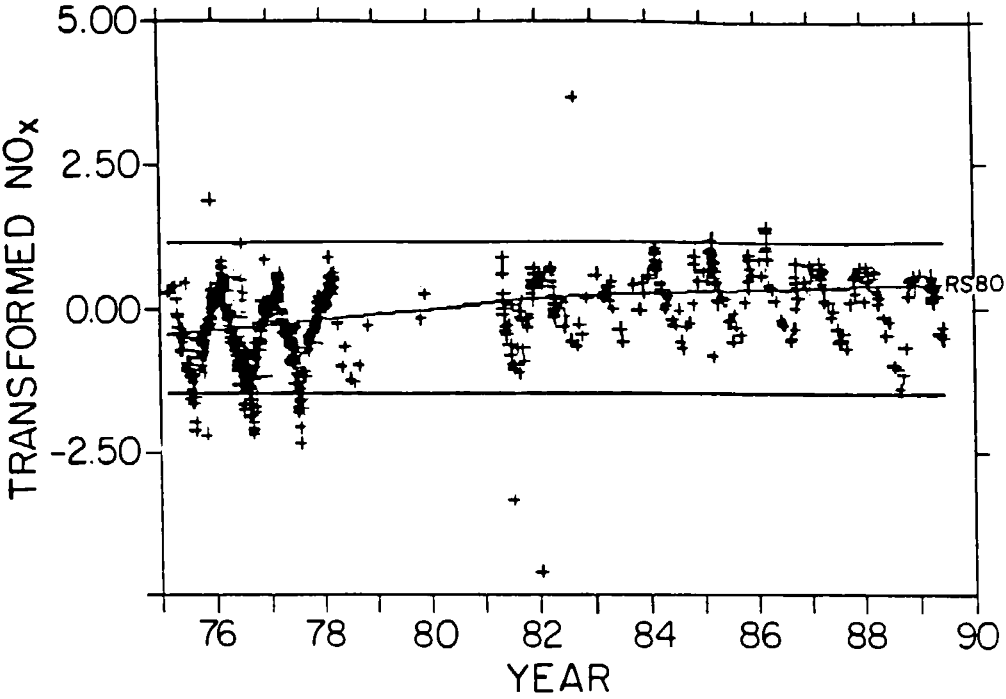


Figure 24.2.1. Graph of logarithmic total nitrates denoted by NO_x (mg/l) against time for the Saugeen River.

In Figure 24.2.1, there is one particularly large observation occurring in 1982. One may wish to examine the records to see if this observation is correct. If the extreme observation were erroneous, one could remove it from the record and thereby not use it in subsequent data analysis. However, the techniques used in the general trend analysis procedure of Section 24.3 are robust or insensitive to outliers. Therefore, this extreme value, and others, are not eliminated from the record.

The line indicating the upward trend through the data is the RLWRS for this time series plot. This robust smooth is referred to in Figure 24.2.1 as RS80 because a value of $f = 0.8$ is employed when plotting the smooth. Hence, the number beside RS is the f value multiplied by 100.

As shown by the dark mass of crosses in the earlier years, more observations were taken at that time. The large gap between many observations, especially in the period from 1979 to 1981, shows that there are time periods during which few measurements were taken and, therefore, the sequence of observations are unequally spaced. The sinusoidal cycle, which is especially pronounced during the first few years, means that the logarithmic total nitrate data are seasonal.

The results of the nonparametric Mann-Kendall trend test (see Section 23.3.2) written below the graph in Figure 24.2.1 confirm that there is an upward trend. This is because the value of the statistic tau calculated using [23.3.5] is positive and the significance level (SL) for this monotonic trend test is close to zero. A small SL (for example, a value less than 0.05) means one should reject the null hypothesis that there is no trend and accept the alternative hypothesis that there is a trend. Alternatively, if the SL level is large and greater than say 0.05, one should accept the null hypothesis that there is no trend. Because the Mann-Kendall test in [23.3.5] or [23.3.1] is designed for employment with nonseasonal data, the result of the test is only a rough indicator for confirming the presence of a monotonic trend in the time series in Figure 24.2.1.

Figure 24.2.2 shows a scatter plot of the logarithmic flows of the Saugeen River (the X variable) against the logarithmic total nitrates (Y variable). The flows are only used for the same times at which the total nitrate measurements are available. Notice that there appears to be a nonlinear function relationship between the flows and the nitrates. To allow the RLWRS to follow this relationship a graph using RS50 is employed. A visual examination of this RLWRS shows that the extreme values do not adversely affect it. The Kendall rank correlation test given at the bottom of Figure 24.2.2 is described in Appendix A23.1. Because tau is positive, there is an upward trend in the scatter plot. The fact that the SL is very small means that the relationship is significant.

24.2.3 Building Regression Models

Overview

Regression analysis constitutes a very general approach to formally modelling statistical data. In fact, regression analysis models can be written in a wide variety of ways and can handle many different situations. When fitting regression models to data sets, one should follow the identification, estimation and diagnostic check stages of model construction, as is done throughout this book for time series models. To illustrate how regression analysis is applied in practice, a case study involving water quality time series is presented.

Lake Erie Water Quality Study

In a particular data analysis study, one should design a specific type of regression model for addressing relevant statistical problems with the data being analyzed. As a brief example of how this is done, consider the statistical data analysis study of water quality time series measured at Long Point Bay in Lake Erie, Ontario, Canada, which is presented in Section 23.5. As summarized in Table 23.5.1, a wide variety of graphical, parametric and nonparametric techniques are utilized for addressing challenging statistical problems. The last item in Table 23.5.1 mentions that regression analysis is employed in the investigation.

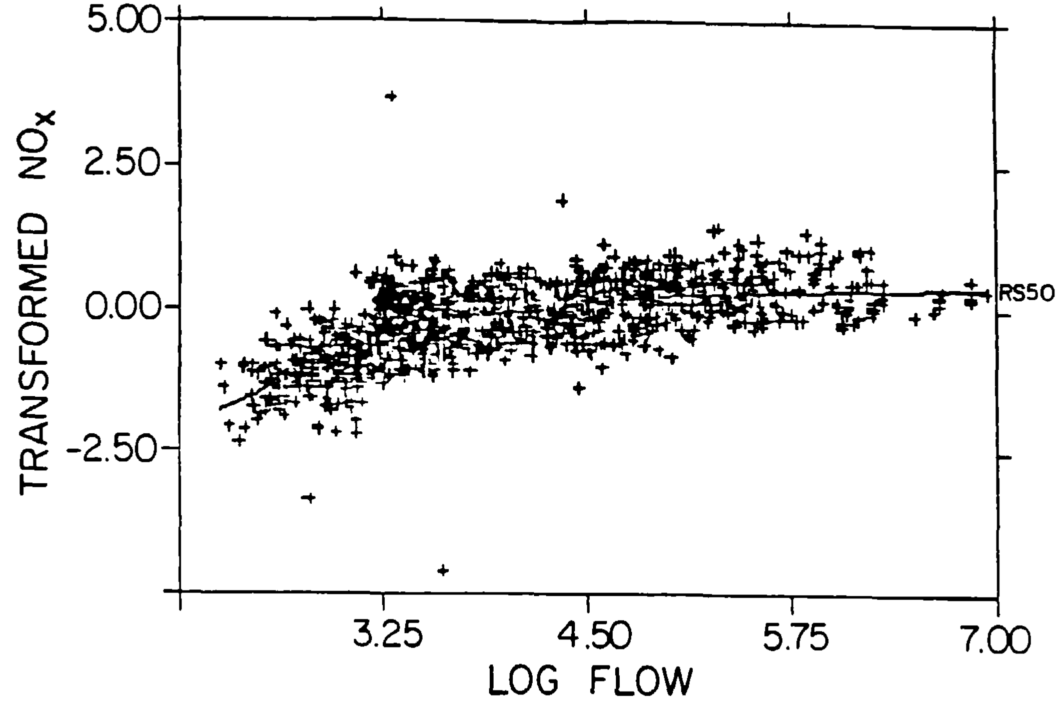


Figure 24.2.2. Scatter plot of logarithmic nitrates against logarithmic flows for the Saugeen River.

Regression Model Design: For the Lake Erie project, a flexible regression model is designed for accomplishing tasks which include determining the best data transformation, ascertaining the components required in a regression model, and estimating both the average monthly and annual values for the series. One can then check for long term trends by examining a smoothed plot of the estimated average annual values.

The most general form of the regression model used in the project is written as

$$y_{ijk}^{(\lambda)} = \mu + \mu_i + \alpha_j + \gamma_{ij} + \beta(x_{ijk} - \bar{x}) + e_{ijk} \qquad [24.2.9]$$

where $i = 1,2,\ldots,n$, denotes the year; $j = 1,2,\ldots,s$, denotes the season when there are s seasons per year (for the monthly Nanticoke data s is usually 9 rather than 12 because often data are not available for the months of January, February and March); $k = 1,2,\ldots,n_{ij}$, denotes the data point in the ith year and jth season for which there is a total of n_{ij} data points; $y_{ijk}^{(\lambda)}$ is the kth data point for a given water quality variable in year i and month j where the λ indicates a Box-Cox transformation (Box and Cox, 1964) which is defined below; μ is the constant term; μ_i is the parameter for the annual effect in the ith year; α_j is the parameter for the seasonal effect in the jth season; γ_{ij} is the interaction term; x_{ijk} is the kth water depth value for the ith year and jth season; $\bar{x}$ is the mean depth across all of the years and seasons; β is the depth parameter; and e_{ijk} is the error for the kth data point in the ith year and jth season and is normally independently distributed with mean zero and variance σ^2. In order for the model to be identifiable, $\mu_n = 0$, $\alpha_s = 0$, $\gamma_{is} = 0$, $i = 1,2,\ldots,n$, and $\gamma_{nj} = 0$, $j = 1,2,\ldots,s$. Also, if $n_{ij} = 0$ then $\gamma_{ij} = 0$. In addition, $\gamma_{ij} = 0$ if $\sum_{j=1}^{s} n_{ij} = 1$ or $\sum_{i=1}^{n} n_{ij} = 1$. If, for example, $n_{ij} \geq 1$ for all i,j then there are $1 + (n-1) + (s-1) + (n-1)(s-1) = ns$ parameters in the model if one ignores the depth parameter, the transformation parameter λ, and the variance of the error.

Box-Cox Transformation: As is also explained in Section 3.4.5, in order to cause the data to be approximately normally distributed and homoscedastic (i.e., have constant variance), the data can be transformed using a Box-Cox transformation (Box and Cox, 1964) which is defined as

$$y_{ijk}^{(\lambda)} = \begin{cases} \lambda^{-1}(y_{ijk}+c)^{\lambda} - 1 & \lambda \neq 0 \\ \ln(y_{ijk}+c) & \lambda = 0 \end{cases} \qquad [24.2.10]$$

where c is a constant which is usually assigned a magnitude which is just large enough to make all entries in the time series positive. Along with maximum likelihood estimates (MLE's) and standard errors (SE's) for the other model parameters in [24.2.9], one can obtain the MLE of λ and its SE for a given data set. Because it is known that MLE's are asymptotically normally distributed, one can obtain the 95% confidence limits for the MLE of a model parameter such as λ.

Automatic Selection Criteria: A wide variety of statistical procedures are available for selecting the best regression model and making sure that certain modelling assumptions regarding the residuals are satisfied. For example, to choose the most appropriate regression model, one can employ automatic selection criteria such as the AIC and BIC defined in [6.3.1] and [6.3.5], respectively.

R^2 Coefficient: A common criterion for assessing the adequacy of fit of a regression model is the square of the multiple correlation coefficient, denoted by R^2. This statistic reflects the proportion of the total variability which is explained by the fitted regression equation. R^2 has a range between 0 and 1 and the higher the value of R^2, the better is the statistical fit. Consequently, when comparing competing regression models, the one with the highest R^2 value is selected.

Whiteness Tests: To test the adequacy of a fitted model, one can check if one or more assumptions underlying the model residuals are satisfied. In particular, one may wish to ascertain if the residuals are random or uncorrelated. The generalized Durbin-Watson test statistic (Wallis, 1972) provides a test of the null hypothesis that there is no autocorrelation in the residuals of a regression model against the alternative hypothesis that there is significant autocorrelation. More specifically, the test statistic is defined as

$$d_k = \sum_{t=k+1}^{n} (\hat{e}_t - \hat{e}_{t-k})^2 / \sum_{t=1}^{n} \hat{e}_t^2 \qquad [24.2.11]$$

where $\hat{e}_t$ is the residual estimated at time t, n is the length of the residual series, and k is a suitably selected positive integer. Based upon the work of Shively et al. (1990) as well as Ansley et al. (1992), Kohn et al. (1993) develop an algorithm for computing the p-value of the test statistic in [24.2.11]. Because of this, one can now conveniently execute an hypothesis test for whiteness using the generalized Durbin Watson test statistic. Additionally, the residual autocorrelation function (RACF) defined at lag k as

$$r_k(\hat{e}) = \sum_{t=k+1}^{n} \hat{e}_t \hat{e}_{t-k} / \sum_{t=1}^{n} \hat{e}_t^2 \qquad [24.2.12]$$

is related to the test statistic in [24.2.11] by the relationship

$$r_k(\hat{e}) \approx 1 - \frac{d_k}{2} \qquad [24.2.13]$$

Kohn et al. (1993) furnish a technique for calculating the p-value for the test statistic in [24.2.13] that allows this statistic to also be used as a whiteness test. With nonseasonal data, for example, one may wish to ascertain if r_1 is significantly different from zero. For a seasonal time series, one may also want to examine $r_k(\hat{e})$ for the case where k is the number of seasons per year or some integer multiple of the seasonal length.

The runs test constitutes another procedure for testing whether or not the residuals are white. The runs test is a simple but often effective test of the null hypothesis that a time series is random. Let M denote the median of a time series $z_1, z_2, \ldots, z_n$. If one replaces each z_t by a + or − according as $z_t \le M$ of $z_t > M$ respectively, then a run is a string of consecutive + or −. The total number of runs, say R, yields a test statistic for randomness. The exact expected number of runs is given by

$$E(R) = 1 + \frac{2n_1 n_2}{n_1 + n_2}, \qquad [24.2.14a]$$

where n_1 is the total number of + and $n_2 = n - n_1$. If there is persistence in the series, the observed number of runs, R, will tend to be less than the expected. On the other hand, for alternating behaviour the number of runs will exceed $E(R)$. The exact variance of R is given by

$$var(R) = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}. \qquad [24.2.14b]$$

Provided that n_1 and n_2 are both greater than 20, the normal approximation can be used to compute the significance level (Swed and Eisenhart, 1943). A two-sided test is used. The Runs Test could be computed about a value other than M but it would have less power. When either $n_1 \le 20$ or $n_2 \le 20$ exact formulae given by Swed and Eisenhart (1943) for the probability function of R are used to compute the exact significance level of a two-sided test.

For the case of the regression model in [24.2.9], one would like to ascertain if the residuals, $\hat{e}_{ijk}$, are random. Consequently, in the above test, the z_t series is replaced by the $\hat{e}_{ijk}$ series. The runs test makes no distributional assumption other than independence. When the residuals of a regression model are not random, this would indicate that the fitted model in inadequate.

Analysis of Variance for the Regression: In order to test the statistical significance of the regression model, an ANOVA (analysis of variance) for regression can be executed. For the model in [24.2.9], the null hypothesis is that all parameters (i.e., the μ_i's, α_j's, γ_{ij}'s, and β) are zero except for the mean which would be given by μ when the other parameters are zero. The alternative hypothesis is that at least one of the parameters, other than the mean, is nonzero.

To calculate the test statistic, various kinds of sums of squares (SS) must be determined. The total SS is

$$SS_{total} = \sum y_{ijk}^2 \qquad [24.2.15]$$

where the $\sum$ refers to summing over all i, j and k. Of course, if the model is fitted to data transformed by a Box-Cox transformation, the SS in [24.2.15] and elsewhere are calculated for the transformed data. The total corrected SS is given by

$$SS_{total\ corrected} = \sum (y_{ijk} - \bar{y})^2 \qquad [24.2.16]$$

where $\bar{y}$ is the overall mean of the y_{ijk}. To calculate the SS of the residuals or the errors, one uses

$$SS_{res} = \sum (y_{ijk} - \hat{y}_{ijk})^2 \qquad [24.2.17]$$

where $\hat{y}_{ijk}$ is the predicted value of y_{ijk} using the fitted regression model in [24.2.9]. The following identity can be used to ascertain the SS for the mean.

$$SS_{mean} = SS_{total} - SS_{total\ corrected} \qquad [24.2.18]$$

Finally, the SS for the regression is given by the identity

$$SS_{reg} = SS_{total} - SS_{mean} - SS_{res} \qquad [24.2.19]$$

To determine the mean square (MS) for a given SS, one divides the SS by the degrees of freedom (DF). The number of degrees of freedom for the regression, denoted by DF_{reg}, is the total number of μ_i, α_j, γ_{ij} and β parameters in [24.2.9] which are not restricted to be zero. If, for instance, $n_{ij} \geq 1$ for all i and j, there are

$$1 + (n-1) + (s-1) + (n-1)(s-1) + 1 = ns + 1$$

degrees of freedom which are due to the mean, μ_i's, α_j's, λ_{ij}'s, and β parameters, respectively. The number of degrees of freedom for the mean correction is simply one while the DF for the SS_{res} is

$$DF_{res} = n - DF_{reg} - 1 \ . \qquad [24.2.20]$$

The MS's for the regression and residuals are given by

$$MS_{reg} = \frac{SS_{reg}}{DF_{reg}} \qquad [24.2.21]$$

and

$$MS_{res} = \frac{SS_{res}}{DF_{res}} , \qquad [24.2.22]$$

respectively.

The test statistic is given by

$$\hat{F} = \frac{MS_{reg}}{MS_{res}} \qquad [24.2.23]$$

which follows a F distribution with DF_{reg} and DF_{res} degrees of freedom. After calculating the test statistic, one can easily determine the SL in order to ascertain whether or not the null hypothesis should be rejected. For this test, it is assumed that the residuals are normally independently distributed with a mean of zero and a constant variance.

Comparing Other Alternative Models: The ANOVA for the regression compares the simplest possible regression model, for which there is only a mean level, to the full model (FM). In general, one may wish to know whether or not a reduced version of the FM, denoted by RM for reduced model, can describe a data set statistically as well as the more complex FM which contains all of the parameters of the simpler RM. The null hypothesis is that the parameters in the FM which are not contained in the RM are all zero. The alternative hypothesis is that at least one of these parameters is nonzero. If, for example, the null hypothesis were accepted based upon the SL of the test statistic, the RM would adequately model the data and the FM would not be required.

Let $\hat{y}_{ijk}$ and $y*_{ijk}$ be the values predicted in the regression equations for the FM and RM models, respectively. The SS of the residuals or errors for the FM and RM are given by

$$SS_{res}(FM) = \sum(y_{ijk} - \hat{y}_{ijk})^2 \qquad [24.2.24]$$

and

$$SS_{res}(RM) = \sum(y_{ijk} - y*_{ijk})^2, \qquad [24.2.25]$$

respectively. The test statistic is then written as

$$\hat{F} = \frac{[SS_{res}(RM) - SS_{res}(FM)]/[DF_{res}(RM) - DF_{res}(FM)]}{SS_{res}(FM)/DF_{res}(FM)} \qquad [24.2.26]$$

where $DF_{res}(FM)$ and $DF_{res}(RM)$ are the numbers of degrees of freedom for the FM and RM, respectively, which are calculated using [24.2.20]. The test statistic in [24.2.26] follows an F distribution with $[DF_{res}(RM)–DF_{res}(FM)]$ and $DF_{res}(FM)$ degrees of freedom. To determine whether or not the null hypothesis should be accepted, the SL for the test statistic can be calculated.

Test for Depth Effect: One may wish to test the hypothesis that an estimated parameter in a fitted regression model is equal to a given constant. If the estimated parameter is $\hat{\beta}_i$, one may wish to test the null hypothesis

$$H_0\text{: } \hat{\beta}_i = \beta_i^0$$

where β_i^0 is a constant selected by the investigator. The alternative hypothesis is

$$H_A\text{: } \hat{\beta}_i \neq \beta_i^0 .$$

The test statistic is

$$t = \frac{\hat{\beta}_i - \beta_i^0}{SE} \qquad [24.2.27]$$

where t follows a student's t distribution on DF_{res} degrees of freedom calculated using [24.2.19], and SE is the standard error of estimation for $\hat{\beta}_i$. After calculating the SL for the statistic in [24.2.26], one can decide whether or not to accept the null hypothesis.

Estimating Monthly Means: After fitting the regression model in [24.2.9] to a given data set, one can obtain estimates for the average monthly values for those months for which at least some measurements were taken. When the data are transformed using the Box-Cox transformation in [24.2.10], the average monthly values are first calculated for the transformed domain. Letting $\hat{v}_{ij}$ stand for the estimate of the average monthly value in year i and month j, then

$$\hat{v}_{ij} = \hat{\mu} + \hat{\mu}_i + \hat{\alpha}_j + \hat{\gamma}_{ij} \qquad [24.2.28]$$

where the definitions for the estimated parameters on the right hand side are given in [24.2.9]. The 95% confidence interval for $\hat{v}_{ij}$ is $\hat{v}_{ij} \pm 1.96SE$. To determine the minimum mean square error (MMSE) estimates of the average monthly means in the untransformed domain, one can use the formulae given by Granger and Newbold (1976) which are discussed in Section 8.2.7. For instance, when $\lambda = 0$ in [24.2.9], the MMSE estimate is

$$\tilde{v}_{ij} = \exp[\hat{v}_{ij} + \frac{1}{2}var(\hat{v}_{ij})] \qquad [24.2.29]$$

To calculate the 95% confidence limits of $\tilde{v}_{ij}$, one can replace $\hat{v}_{ij}$ by $\hat{v}_{ij} + 1.96SE$ and $\hat{v}_{ij} - 1.96SE$ in order to determine the upper and lower limits, respectively, for the transformed domain.

Estimating Annual Means: By letting $\hat{v}_i$ represent the estimate of the average annual value for year i in the transformed domain, the annual mean for year i can be calculated using

$$\hat{v}_i = \frac{1}{s}\sum_{j=1}^{s} \hat{v}_{ij} \qquad [24.2.30]$$

where s is the number of seasons for which the v_{ij} are estimated. The variance of $\hat{v}_i$ is determined as

$$var(\hat{v}_i) = \frac{1}{s^2}\sum_j \sum_h cov(v_{ij} v_{ih}) \qquad [24.2.31]$$

The 95% confidence limits for $\hat{v}_i$ are $\hat{v}_i \pm 1.96SE$ where SE is the square root of the variance in [24.2.28]. To determine the MMSE estimate, $\tilde{v}_i$, of the average annual value for year i in the untransformed domain one can employ the formulae of Granger and Newbold (1976). The 95% confidence limits for $\tilde{v}_i$ are found by taking the inverse Box-Cox transformation of the 95% confidence limits in the transformed domain. To determine visually if there is a long term trend, one can plot a RLWRS or other kind of smoothed curve through the estimated annual time series. One could also produce a separate graph of the Tukey smooth for the annual values described in Section 22.3.5.

24.3 TREND ANALYSIS METHODOLOGY FOR WATER QUALITY TIME SERIES MEASURED IN RIVERS

24.3.1 Introduction

The collection and analysis of water quality time series are of great import in many regions throughout the world, especially in highly populated and industrialized areas. For example, the Ministry of the Environment within the Canadian province of Ontario operates a spatially and temporally extensive sampling network called the Provincial Water Quality Monitoring Network or simply PWQMN. Approximately one sample per month is collected at over 700 sites, which may be analyzed for up to 60 water quality indicator parameters. In fact, the PWQMN is one of the world's largest water quality sampling networks falling under the umbrella of a single political jurisdiction. Large sums of money are being spent on collecting substantial data through the PWQMN. To make this data meaningful and useful, they must be properly summarized and analyzed. The Ministry of the Environment, as well as many other organizations, are especially interested in detecting and modelling historical trends in PWQMN data. Trend analyses are required for alerting authorities about water quality degradation so that appropriate corrective action can be taken and for evaluating the performance of pollution abatement schemes.

The purpose of this section is to present a general methodology for analyzing trends in water quality time series measured in rivers. When checking for the presence of a trend in a water quality time series, the methodology properly takes into account the effects of riverflows and seasonality upon the water quality observations. Furthermore, the methodology can be used with messy water quality data (see Section 23.1) which may possess undesirable characteristics such as having outliers, non-normality and missing values.

To design the steps presented in the methodology, the authors examined a wide variety of PWQMN water quality time series measured in the Saugeen and Grand Rivers in Southern Ontario, Canada. Based upon the many types of trend analysis problems that arose when analyzing the data, a systematic procedure for studying the time series was developed. Because unforeseen problems were discovered as different kinds of water quality data were analyzed, the trend analysis algorithm was built and improved in an iterative fashion. The final product is a comprehensive and flexible trend analysis methodology for carrying out systematic trend studies of water quality time series.

Within the steps in the methodology, specific graphical, parametric and nonparametric techniques described in Part X of this book are utilized. Although the authors found these techniques to be sufficient for handling all the trend analysis problems they encountered, practitioners and researchers may wish to employ additional specific methods at certain steps in the algorithm. For instance, when looking for basic characteristics of the data by examining graphs of the data, some people may wish to use graphical methods beyond those presented in Section 22.3. Whatever the case, the main steps in the algorithm will remain the same.

In the next section, the steps in the trend analysis methodology are presented and practical applications are employed to demonstrate how the methodology can easily be applied in practice. Although the authors actually applied their methodology to eight PWQMN water quality series plus one waterflow sequence measured at two locations in Southern Ontario, only some representative results are given to explain how the methodology works. Finally, an earlier version of research appearing in this section is provided by McLeod et al. (1991).

24.3.2 Methodology Description

Overview

The methodology given below is valid for use with messy water quality series measured in rivers. In the description, it is assumed that one is ultimately wishing to detect trends at the monthly level. Nonetheless, the methodology can be easily converted for use with other seasonal levels such as quarter-yearly or weekly. To explain the procedure, the NO_x and riverflow data for the Saugeen River at Burgoyne, Ontario, Canada, are employed.

The overall trend analysis study is divided into the two main categories of Graphical Studies and Trend Tests. These two groupings reflect the idea of exploratory and confirmatory data analyses referred to in Sections 1.2.4, 22.1, 23.1 and 24.1. Within the category of Graphical Studies, the following three versions of the water quality series are examined first for trends:

1. *Raw or unadjusted water quality time series*: The given series may be transformed by the Box-Cox transformation in [3.4.30] or [24.2.10] in an attempt to make a non-normal time series become approximately normally distributed (see discussion in Section 3.4.5 and 24.2.3).

2. *Flow-adjusted water quality time series*: This is the time series for which effects of flow upon water quality are removed, as explained in detail below. As mentioned in Section 24.1 just after [24.1.1], in certain situations one may wish to use a covariate series other than flow to adjust the water quality series. If this is the case, replace the word flow by the name of the covariate in the general methodology described in this section.

3. *Detrended-flow-adjusted water quality time series*: After removing trends from the water quantity time series, the influences of flow upon the water quality time series are eliminated in order to produce the detrended-flow-adjusted water data.

Following this, the three average monthly versions of the above three kinds of data are studied using graphical procedures. As noted at the start of this section, one can easily use a seasonal time scale other than monthly. If this is required, replace the word monthly by the designated seasonal category in the description of the methodology.

4. *Mean monthly unadjusted water quality time series.*

5. *Mean monthly flow-adjusted water quality time series.*

6. *Mean monthly detrended-flow-adjusted water quality time series.*

The manner in which these series are calculated is explained below. The main graphical procedure used to examine the six types of water quality data are a trace or time series plot (Section 22.3.2) along with a smoothed curve (called RLWRS in Section 24.2.2) through the plotted data. Finally, under the category of trend tests, the above three types of monthly water quality data are analyzed using trend tests from Chapter 23. Of particular importance is the Spearman partial rank correlation test described in Section 23.3.6 which works extremely well with seasonal data.

The overall trend analysis methodology is summarized in Table 24.3.1. Specific details are presented in Table 24.3.2 for carrying out the Spearman partial rank correlation mentioned opposite d in Table 24.3.1. The steps in the methodology are now explained in detail using the total nitrate (i.e. NO_x) data measured in the Saugeen River at Burgoyne.

Table 24.3.1. Trend analysis methodology for use with water quality time series measured in rivers.

TREND ANALYSIS METHODOLOGY

GRAPHICAL TREND STUDIES

Examine traces along with smoothed curves (i.e. RLWRS's) for the following data sets:

Given Data:

1. Unadjusted water quality time series.
2. Flow-adjusted water quality time series.
3. Detrended-flow-adjusted water quality time series.

Mean Monthly Data:

4. Mean monthly unadjusted water quality time series.
5. Mean monthly flow-adjusted water quality time series.
6. Mean monthly detrended-flow-adjusted water quality time series.

TREND TESTS

For the three mean monthly data sets (i.e. 4, 5 and 6), the following trend tests are carried out:

a. Mann-Kendall (Section 23.3.2).

b. Spearman's rho (Section 23.3.6).

c. Seasonal Mann-Kendall (Section 23.3.2).

d. Spearman partial rank correlation when partialling out seasonality (Section 23.3.6).

To test for seasonality, the Kruskal-Wallis test (Appendix A23.3) and box and whisker graphs (Section 22.3.3) can be used. The tests under c and d are designed for use with seasonal data.

Table 24.3.2. Algorithm for the Spearman partial rank correlation trend test when partialling out seasonality.

ALGORITHM

1) X_t is one of the three monthly series given under 4, 5 and 6 in Table 24.3.1.

2) Test for the presence of seasonality in X_t using

   a. Box and whisker graphs (Section 22.3.3).

   b. Kruskall-Wallis test (Appendix A23.3).

3) If seasonality is not found, use the ordinary Mann-Kendall trend test (Section 23.3.2).

4) When seasonality is present, carry out the Spearman partial rank correlation test of Section 23.3.6 where:

   a. X_t is the series from 1).

   b. $Y_t = t$ where t is the time of the observation.

   c. Z_t is obtained from the ranking of the seasonal effects in the Kruskall-Wallis test (Appendix A23.3).

Graphical Trend Studies

Given Data:

1. Unadjusted water quality time series: As indicated in Table 24.3.1, the first step is to examine a trace along with a RLWRS of the given unadjusted data or the data transformed by a Box-Cox transformation in [3.4.30] or [24.2.10]. A common transformation is to take natural logarithms of the data (i.e. $\lambda = 0$ in [3.4.30]). A graph of the logarithmic total nitrates (NO_x) for the Saugeen River is displayed in Figure 24.2.1. One can easily see the increasing trend over time traced by the RLWRS. Additionally, the results of the Mann-Kendall test at the bottom of the graph also confirm the presence of a trend. However, the SL may not be meaningful because of the high degree of correlation in the data caused by frequent sampling at particular time periods, especially from 1976 to 1978. Moreover, there is also strong seasonality in this data which the Mann-Kendall test cannot properly take into account.

2. Flow-adjusted water quality time series: The question arises as to whether or not a given water quality variable is dependent upon flow. Figure 24.2.2 displays a scatter plot of the logarithmic NO_x series against logarithmic flows. Each flow value is plotted for exactly the same time at which the corresponding NO_x observation is made. As shown by the RLWRS, there is an obvious dependency between NO_x and flow. The value of the Kendall rank correlation test statistic (Appendix A23.1) listed at the bottom of Figure 24.2.2 for the data plotted in this figure, is also significantly large and, therefore, confirms this finding.

Each sample flow value in Figure 24.2.2 is the average daily value for the day on which the corresponding NO_x value was collected. For relatively large rivers such as the Saugeen and

Grand, the discrepancy between instantaneous flow for the exact point in time at which the water quality sample is collected and the mean daily flow, is generally negligible.

Because the water quality values and flows are dependent, one would like to see if a trend is present in the logarithmic water quality time series after the flow effects are removed. To accomplish this, one can examine the residuals of the RLWRS in Figure 24.2.2 . To calculate the residual for each plotted point in Figure 24.2.2, one subtracts the value of the RLWRS at that point. For convenience, a RLWRS using RS50 is used when calculating the residuals. This series is called the flow-adjusted water quality time series. Other approaches for obtaining flow-adjusted water quality time series are discussed in Section 23.3.5.

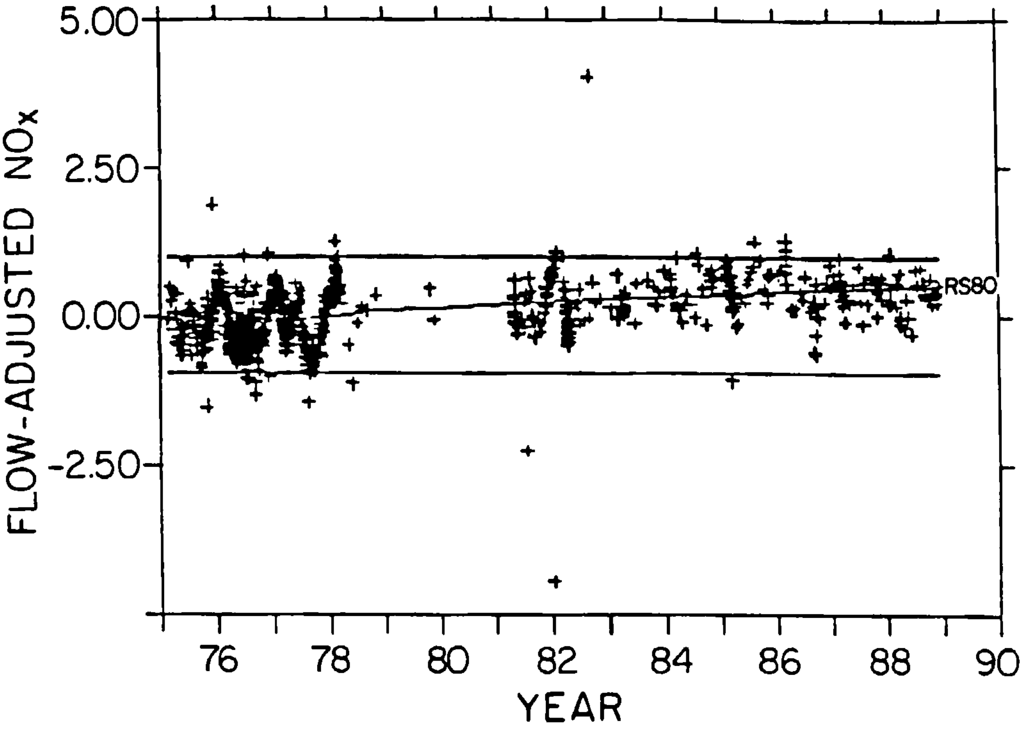


Figure 24.3.1. Graph of the flow-adjusted NO_x series against time for the Saugeen River.

Figure 24.3.1 presents a trace of the flow-adjusted NO_x data for the Saugeen River. Notice that even after flow effects are removed, the RLWRS still shows an obvious upward trend over time.

3. **Detrended-flow-adjusted water quality time series**: If a trend were present in the flow or covariate series, one would want to remove this trend and then subsequently adjust the water quality time series for detrended flow. In this step, a flexible procedure for obtaining a detrended-flow-adjusted water quality series is described.

Suppose, for now, that the NO_x series really does not have a trend. Flows may cause a trend to appear in the series due to one or both of the following two reasons. First, there may be a real trend in the flows which also causes a trend in the NO_x series. Second, thę sampling bias of the flows may cause a trend. Recall that in Figure 24.2.2, each flow is plotted for exactly the same time as the corresponding NO_x observation so that many of the flow observations are not used when producing the flow-adjusted NO_x series. When the complete series of logarithmic Saugeen flows are plotted against time, no trend is present. However, when the logarithmic flows are plotted against time for exactly the same times at which the NO_x values are measured, Figure 24.3.2 shows that the sampling bias has created an obvious trend. To remove the trend from the logarithmic flow series in Figure 24.3.2, one can subtract the RLWRS value from the logarithmic flow series at each time point for which an NO_x observation is available. This residual series is determined for a RLWRS using RS50 to obtain the detrended logarithmic flow series.

Figure 24.3.3 displays a scatter plot of the logarithmic NO_x series against the detrended logarithmic flows. Notice that there is still a dependence between the two series even after the trend due to sampling bias is removed from the logarithmic flows. To obtain the detrended-flow-adjusted NO_x series, one uses the residuals of the RLWRS for the case when RS50 is used to determine the smooth.

For the Saugeen River data, the complete flow record possesses no trend. Because of this, one can state that the trend in the partial flow record plotted in Figure 24.3.2 is due to sampling bias. If the complete record of flows contained a trend, then a trend in the partial flow record (i.e. those flows occurring on the same days at which the water quality samples were collected) would be due to an actual trend in the flows and perhaps also sampling bias.

Figure 24.3.4 shows a graph of the detrended-flow-adjusted data against time. Both the RLWRS and the Mann-Kendall trend test result in this figure show that there is a trend.

Mean Monthly Data

4. **Mean monthly unadjusted water quality series**: To ascertain the behaviour of the NO_x series at the monthly level, one can examine graphs of the average monthly series. One should keep in mind that the term "average" refers to calculating a mean which may be determined from only a few observations in a given month, each of which is collected in a 10 to 15 second time interval (see discussion on missing values in Section 24.2.1). Figure 24.3.5 shows a graph of the logarithms of the mean monthly NO_x series for the Saugeen River. The RLWRS shows that there is an increasing trend over time. Additionally, there are some months for which no observations are available.

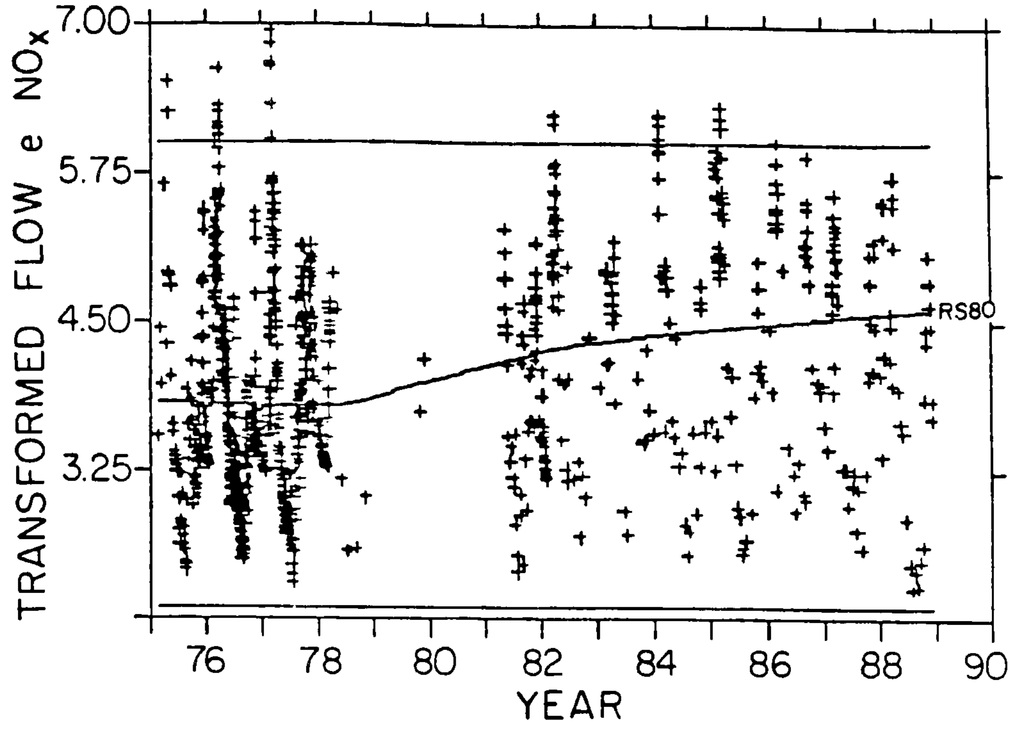


Figure 24.3.2. Logarithmic Saugeen flows against time plotted at exactly the same times at which NO_x observations are available.

5. **Mean monthly flow-adjusted water quality time series**: Except for the fact that monthly values are used, the logarithmic mean monthly flow-adjusted NO_x series is calculated in exactly the same way as the logarithmic flow-adjusted water quality time series for the given data under item 2. Hence, one determines the residuals of the RLWRS using RS50 fitted to a scatter plot of logarithmic mean monthly NO_x series against the logarithmic mean monthly flows. Figure 24.3.6 displays the graph of the logarithmic mean monthly flow-adjusted water quality series against time, for which there is a striking linear upward trend.

6. **Mean monthly detrended-flow-adjusted water quality time series**: To eliminate the effects of trends and/or sampling bias in the monthly flows upon the average monthly NO_x series, one can calculate the average monthly detrended-flow-adjusted data. The same procedure followed for determining the flows under item 3 for the given data is also employed here. First, one fits a RLWRS using RS50 to a graph of the logarithmic mean monthly flows against time where the flow observations are only used for months for which NO_x values are available. The

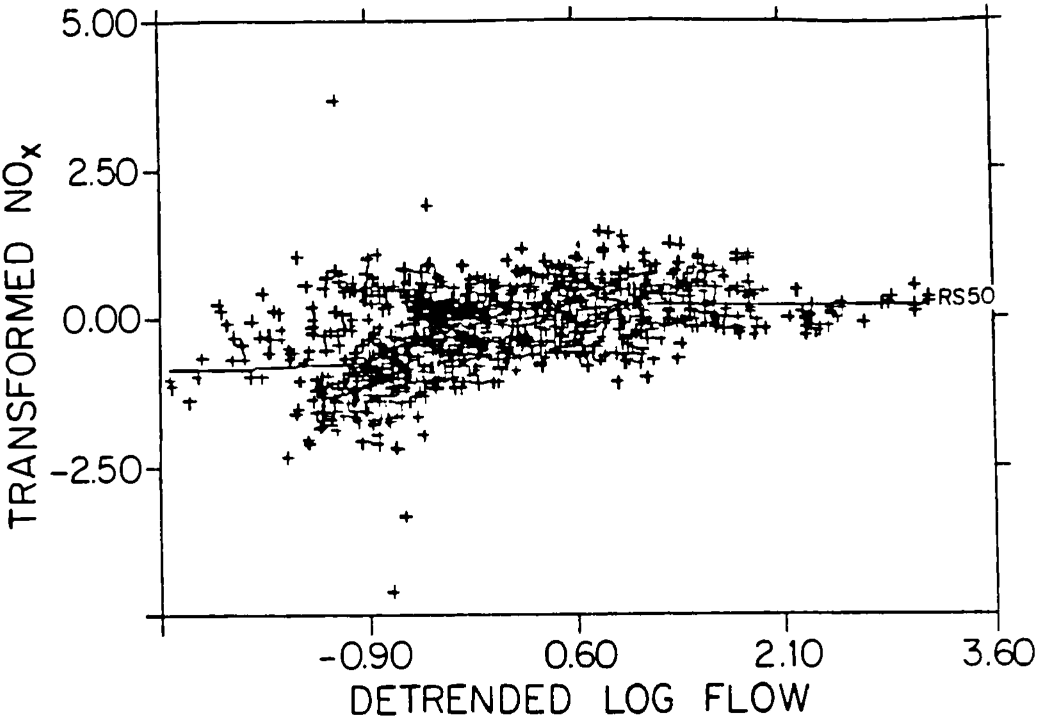


Figure 24.3.3. Scatter plot of the logarithmic NO_x data against detrended logarithmic flows for the Saugeen River.

residuals of the smooth form the logarithmic mean monthly detrended flow series. Second, a RLWRS smooth using RS50 is fitted to a scatter plot of the logarithmic mean monthly NO_x series against the logarithmic mean monthly detrended flow series. The residuals of this smooth constitute the logarithmic average monthly detrended-flow-adjusted NO_x series.

Figure 24.3.7 displays a plot of the logarithmic average monthly detrended-flow-adjusted time series against time. The RLWRS clearly reveals the increasing trend present in this data.

Trend Tests

The four trend tests listed in the bottom half of Table 24.3.1 are applied separately to each of the three types of mean monthly series. For all three series, the four trend tests give exactly the same results. Because of this, representative findings are presented for only one of the three monthly series for explanation purposes in this section.

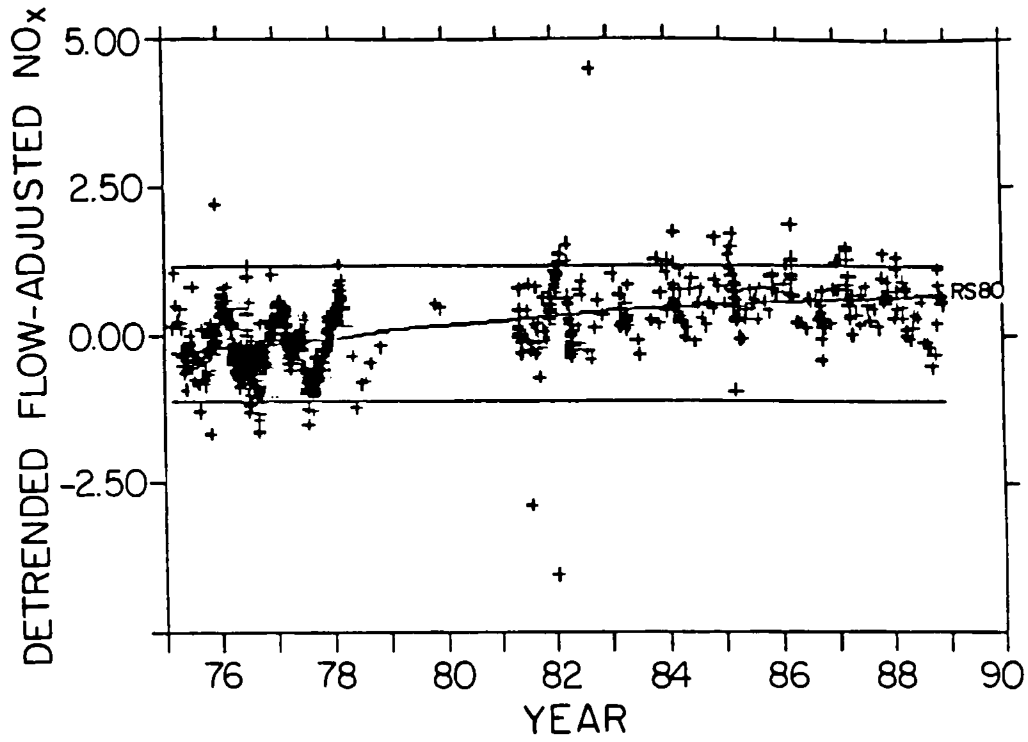


Figure 24.3.4. Graph of detrended-flow-adjusted logarithmic NO_x observations against time for the Saugeen River.

Consider the fifth series which is the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River. Table 24.3.3 presents the results for the four trend tests for this series while Table 24.3.4 gives the average rank value and rank found for each of the twelve seasons used in the Kruskal-Wallis seasonality test. Finally, Figure 24.3.8 displays the box and whisker graph for each of the 12 months for the series being considered. The reader can refer to Section 23.3 for descriptions of the statistical trend tests listed in Table 24.3.3, Appendix A23.3 for a presentation of the Kruskal-Wallis test used in Table 24.3.4, and to Section 22.3.3 for an explanation of box and whisker graphs.

Each of the four trend tests findings in Table 24.3.3 demonstrate that there is a significant trend. In particular, notice that the significance levels are very close to zero for the Mann-Kendall (Section 23.3.2), Spearman's rho (Section 23.3.6), seasonal Mann-Kendall (Section 23.3.2) and the Spearman partial rank correlation (Section 23.3.6) trend tests.

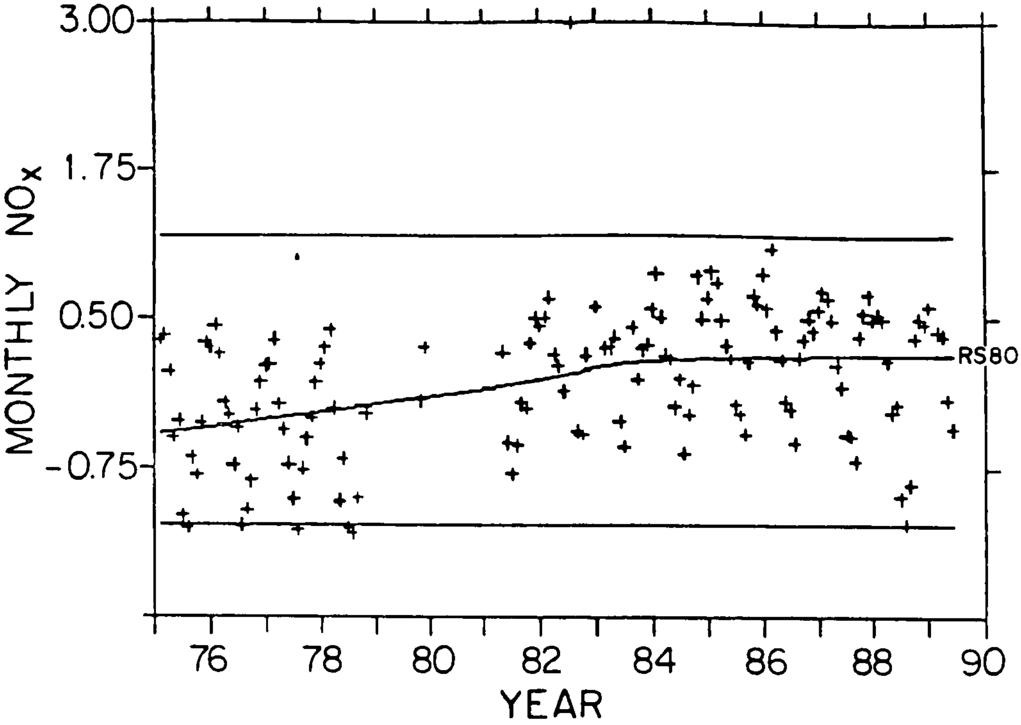


Figure 24.3.5. Graph of the logarithmic average monthly series against time for the Saugeen River.

The test statistic in Table 24.3.4 for the Kruskall-Wallis test has a value of 47.519 and a significance level close to zero. Hence, there is seasonality present in the series. One can also see the cyclic pattern caused by seasonality in the box and whisker graphs for this NO$_x$ series in Figure 24.3.8. Notice, in particular how the median levels change from month to month.

Figure 24.3.8 is an example of what is called a *notched box-and-whisker graph* in Section 22.3.3. The notches on both sides of a box can be used to ascertain if the median in one month is significantly different from another. In particular, when comparing two months, if the median bar in one month overlaps with the notch in the other, and vice versa, then one can argue that the medians for these two months are not significantly different from one another at the 5% significance level. When there are not many data points used to determine a box and whisker plot for a given season, any peculiarities in the plot should be cautiously considered. In Figure 24.3.8, the varying median levels across the months show that the data are seasonal. Notice also for some months that a notch for the mean may extend above an upper hinge or below a lower hinge.

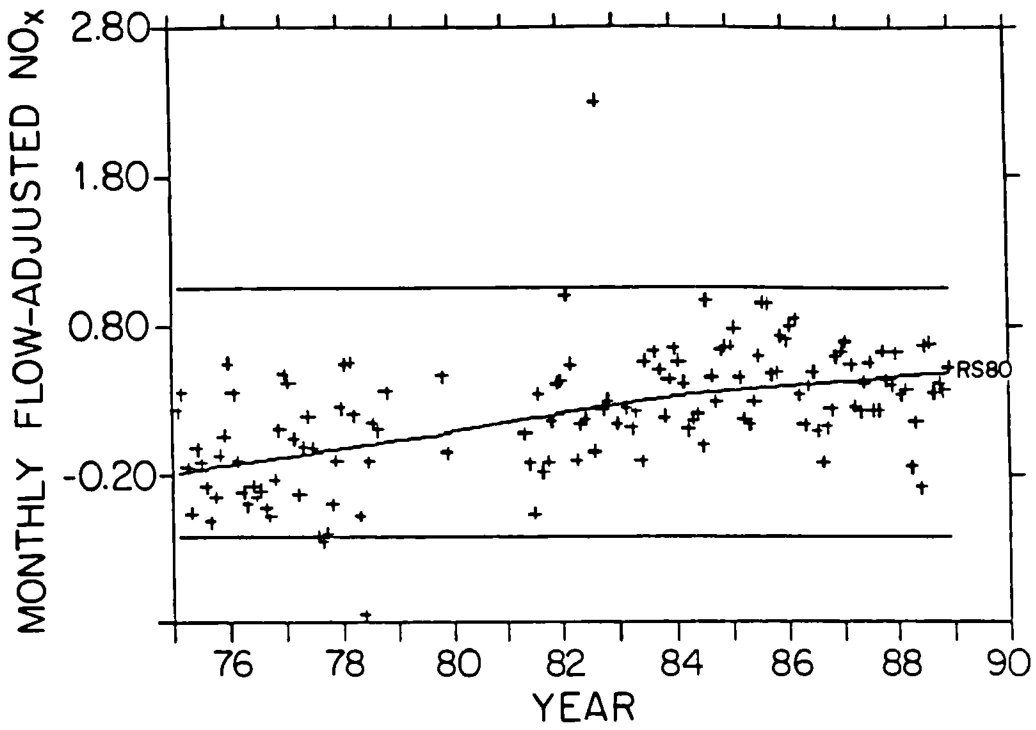


Figure 24.3.6. Graph of the logarithmic mean monthly flow-adjusted NO_x series against time for the Saugeen River.

Because of the importance of the *Spearman partial rank correlation test* for detecting trends in seasonal data, consider the results of this test in more detail. The algorithm for this test is given in Table 24.3.2. In this case, the X_t is the fifth series which is the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River. Under Step 2 of the algorithm in Table 24.3.2, the Kruskall-Wallis test result in Table 24.3.4 as well as the box and whisker graphs of Figure 24.3.8 demonstrate that the data are seasonal. In Table 24.3.4, the seasons are ranked from smallest to largest according to the average rank values for the months. The monthly medians in Figure 24.3.8 can also be compared to obtain the same rankings. Following Step 4 of the algorithm in Table 24.3.2, one lets the Y_t series in the Spearman test be time t while Z_t consists of the seasonal ranks where each month across the years is always given the same rank. By substituting into [23.3.35], one can obtain the Spearman test statistic which has a value of 0.572. Because the SL is almost zero, there is a significant trend over time in the data when the effects of seasonality are partialled out. Since the test statistic is positive, the trend is increasing over time.

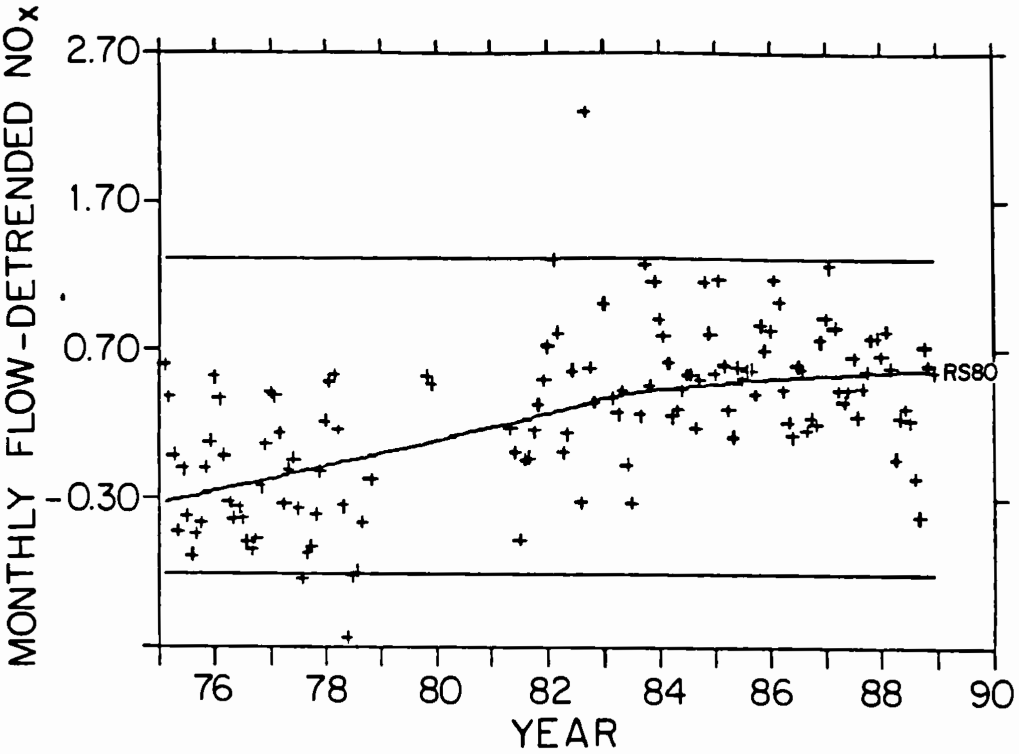


Figure 24.3.7. Graph of the logarithmic average monthly detrended-flow-adjusted data for the Saugeen River.

Table 24.3.3. Trend test results for the logarithmic mean monthly flow-adjusted NO$_x$ series for the Saugeen River at Burgoyne.

Trend Tests	Test Statistics	Significance Levels
Mann-Kendall	0.368	0.000
Spearman's Rho	0.542	0.000
Seasonal Mann-Kendall	314	0.000
Spearman Partial Rank Correlation	0.572	0.000

Table 24.3.4. Average ranks and ranks from the Kruskal-Wallis analyses for the 12 months of the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River at Burgoyne.

Months	Sample Sizes	Average Rank Values	Ranks
1	10	89.50	6
2	10	107.70	7
3	11	107.40	7
4	11	92.27	6
5	12	42.00	2
6	12	37.17	1
7	11	46.18	2
8	11	60.82	4
9	12	68.33	5
10	11	54.36	3
11	13	57.92	3
12	11	64.91	4

Test statistic = 47.519 Significance level = 0.000

24.3.3 Summary

A flexible and comprehensive trend analysis methodology is now available for carrying out a systematic study for detecting and modelling trends in water quality series measured in rivers. As summarized in Table 24.3.1, the two main components to the methodology are the Graphical Trend Studies and the Trend Tests. Of particular import and usefulness for analyzing trends in seasonal water quality data is the Spearman partial rank correlation test of Section 23.3.6. When using this test for detecting a trend in a seasonal series for which seasonality is partialled out, the Spearman algorithm of Table 24.3.2 can be utilized. Finally, the overall methodology contains procedures for accounting for the effects of flow upon a given water quality variable.

At various locations in Section 24.3.2, it is noted that one could, if required, use additional graphs and trend tests within the overall trend analysis methodology of Table 24.3.1. For instance, one may wish to employ the adjusted variable Kendall trend test proposed by Alley (1988). However, the authors found that the specific exploratory and confirmatory techniques presented in Part X, readily handled all the situations that arose when examining the water quality series from the Saugeen and Grand Rivers in Southern Ontario.

Another approach to the trend analysis would be to add a deseasonalization step either before or after Step 2 in Table 24.3.1. However, this procedure is not followed here for a number of reasons. First, adjusting the water quality series for flow or some other covariate series may also remove some seasonality. Secondly, an efficient procedure for removing seasonality from a wide variety of messy water quality time series may be very difficult to design. Finally, when seasonality is present in the data sets numbered 4 to 6, a seasonal trend test can be used for checking for the presence of trends in seasonal data (trend tests under c and d in Table 24.3.1).

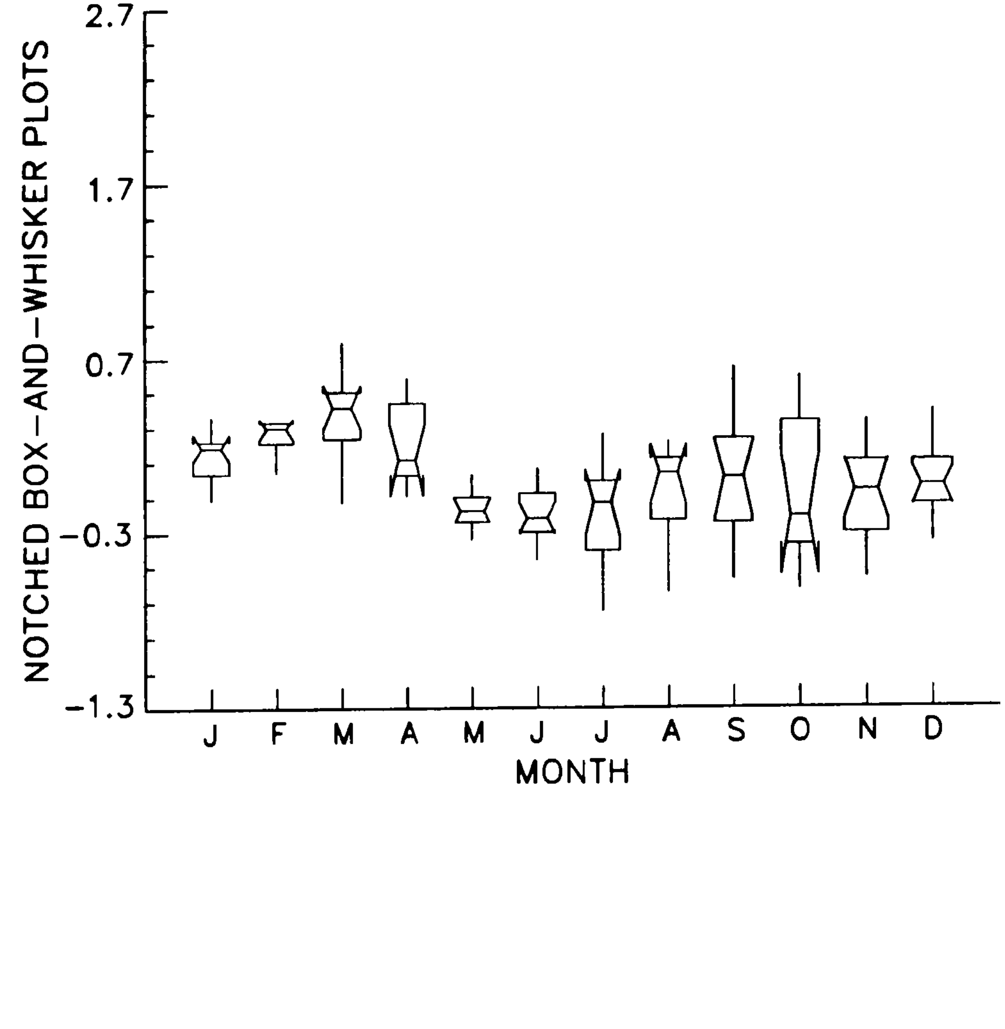


Figure 24.3.8. Box and whisker graphs for the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River.

As demonstrated by the application to the total nitrates data for the Saugeen River, the methodology of Section 24.3.2 works well in practice. In Table 24.3.5, a summary of the findings for the NO_x data is presented. For both the graphical trend studies and the trend tests, trends are always detected in the versions of the NO_x series that are examined.

Table 24.3.5. Summary of the trend analysis results for the total nitrates data measured in the Saugeen River at Burgoyne.

Data Transformation: Logarithmic

Seasonality: Very strong

Flow-Concentration Relationship: Positive relationship at low flow with relatively constant relationship at higher flow.

Outliers: A few (see Figure 24.2.1). Keep in mind that all techniques used in Table 24.3.1 are robust to outliers.

Trend: All tests indicate a significant trend. Examination of the trace plots suggest that it is largely due to a difference in levels in the data from 1975-1978 and 1982-1989.

Other: Very little data over the period 1979-1981.

As reported elsewhere by McLeod et al. (1991), the authors have used the general trend analysis methodology of Section 24.3.2 with many other water quality time series. In particular, they applied the methodology to the eight PWQMN water quality series (mg/l) listed in Table 24.3.6, as well as riverflow series (m^3/s), for the Saugeen River at Burgoyne and also the Grand River at Dunnville, Ontario. The data at these two sites were selected for study because flow biased monitoring was used. This means that more samples were collected at high flows for the purpose of mass-discharge estimation. Hence, the data collected at these high frequency monitoring sites should contain greater relevant information and their analyses should provide insight into how to analyze both highly monitored and less frequently monitored sites. Representative trend analysis results for the NO_x series for the Saugeen River are employed in Section 24.3.2 for explaining how to apply the methodology.

24.4 CONCLUSIONS

Regression analysis provides a set of flexible statistical tools which can be useful in environmental impact assessment studies as exploratory and confirmatory data analysis methods. A particularly flexible smoothing technique which can be employed as an exploratory data analysis method, for tracing trends in a time series, is the RLWRS of Section 24.2.2. Lewis and Stevens (1991) provide another informative regression approach for drawing a trend curve through a time series. As explained in Section 24.3.2, the RLWRS can also be utilized for removing trends from a series as well as dependent relationships between two series. In Section 24.2.3, a specific case study is used for explaining how a regression analysis model can be

Table 24.3.6. Water quality variables measured in the Saugeen River at Burgoyne and the Grand River at Dunnville, Ontario, Canada.

Water Quality Variables
Ammonia Nitrogen
Total Kjeldahl Nitrogen
Total Nitrates (NO_x)
Filtered Reactive Phosphorus
Total Phosphorus
Suspended Solids
Alkalinity
Conductivity

designed as a confirmatory data analysis technique.

A flexible and comprehensive trend analysis methodology is now available for detecting trends in water quality data measured in rivers. As described in Section 24.3.2, the trend analysis procedure consists of the two main stages of graphical trend studies and trend tests. Specific graphical techniques and statistical trend tests that can be employed in the two main stages are described in detail in Sections 22.3 and Chapter 23, respectively. A particularly powerful trend test for use with seasonal water quality data is the Spearman partial rank correlation test given in Section 23.3.6. Table 24.3.2 presents an algorithm for applying the Spearman partial rank correlation trend test when partialling out seasonality. Application of the trend analysis methodology to water quality series measured in the Saugeen and Grand Rivers demonstrates that the procedure works well in practice.

The overall trend analysis methodology outlined in Table 24.3.1 contains many original developments in environmental impact assessment. Firstly, the RLWRS of Section 24.2.2 is used for calculating flow-adjusted and detrended-flow-adjusted water quality series. Secondly, by employing the detrended-flow-adjusted water quality procedure one can eliminate sampling bias when the entire riverflow series does not possess a trend. As a third contribution, the methodology suggests testing for the presence of seasonality before applying a seasonal trend test such as the seasonal Mann-Kendall test of Section 23.3.2. Simulation studies show that when the data are not seasonal, the seasonal Mann-Kendall test is not as powerful as the Mann-Kendall trend test. Fourthly, the Spearman partial rank correlation test (Section 23.3.6) when partialling out seasonality provides a powerful test for use with seasonal water quality time series. As noted in Section 23.3.6, the Kendall partial rank correlation test cannot be used for this purpose since the distribution of the test statistic is unknown and probably analytically intractable.

PROBLEMS

24.1 Beyond the literature cited in this chapter, locate three other references in which regression analysis is applied to water resources or environmental engineering problems. Briefly explain the purpose of and approach for using regression analysis in each of the papers. Moreover, outline the benefits and drawbacks of employing regression analysis for each of the case studies and suggest how improvements could be made.

24.2 A general approach for trend analysis is given in [24.1.1]. Find a paper in the environmental engineering literature not cited in Chapter 24 in which regression analysis is employed for describing the situation in [24.1.1]. Summarize how the regression analysis is used and explain the advantages and disadvantages of employing the regression procedure as given by the authors of the paper.

24.3 Alley (1988) and also Smith and Rose (1991) describe two basic ways in which regression analysis can be employed in trend assessment. Explain how each of these procedures is carried out and compare their relative strengths and weaknesses.

24.4 Many authors, including Pearson (1897), Huff (1954), Good (1959, 1978), Benson (1965), Wong (1979), Kenny (1982), Wong and DeCoursey (1986), Kite (1989), and Kronmal (1993), discuss statistical fallacies including those arising from the use and abuse of regression analysis. By referring to appropriate literature, clearly explain how spurious correlations and other problems can take place when regression analysis is improperly utilized and how these problems can be overcome.

24.5 Beauchamp et al. (1989) compare regression and time series methods for synthesizing missing streamflow records. After summarizing how they carry out their study, comment upon their findings.

24.6 Outline how the approach of Esterby and El-Shaarawi (1981a,b) and El-Shaarawi and Esterby (1982) works for detecting a point of change in a regression model and estimating the magnitude of the change.

24.7 Concepts from fuzzy set theory have now been incorporated into regression analysis. By referring to the appropriate literature, outline the theory and practice of fuzzy regression analysis and discuss the dividends that can be gained by employing this approach. Describe a hydrological application of fuzzy regression analysis, including a discussion of the insights that are found about the problem being studied.

24.8 As pointed out in Section 24.2.2, Cleveland et al. (1990) have developed a seasonal-trend decomposition procedure based upon the RLWRS (Cleveland, 1979). Outline the main steps in this technique and discuss its advantages and drawbacks. Apply the procedure to a seasonal time series which is of interest to you and be sure to mention any insights which you gain.

24.9 Select two nonseasonal time series between which you feel a meaningful relationship may exist. Plot the RLWRS of Section 24.2.2 on a scatter plot of these two series. Experiment with various values of the smoothing variable, f, where

$0 < f \le 1$. Discuss the insights that are provided by the graph.

24.10 Carry out the instructions of Problem 24.9 for two seasonal time series.

24.11 Choose a nonseasonal time series that you think may contain a trend. On a time series plot of the series against time, draw the RLWRS. Comment upon the behaviour of any trend that you find in your data set.

24.12 Execute the instructions of Problem 24.11 for the case of a seasonal time series.

24.13 The sample autocorrelation function (ACF), r_k, in [2.5.9] provides a means for quantifying the linear dependence between values in a time series separated by k time lags. To visualize dependence within a single time series, z_t, one can draw a scatter plot of z_t against z_{t-k} along with a RLWRS. Select a nonseasonal time series which is of interest to you and produce a scatter plot and RLWRS of z_t versus z_{t-k} for $k = 1,2, \ldots ,7$. Comment upon the type of dependence that you can visually detect in each of the scatter plots. Also, calculate r_k in [2.5.9] for $k = 1,2, \ldots ,7$, and compare these results to the visual findings.

24.14 In Section 24.2.3, a specific regression model is designed for modelling water quality time series measured in a lake. Locate a paper in the environmental engineering literature in which the authors employ regression analysis. By using equations when necessary, clearly explain how the authors design, calibrate and check the residual assumptions of their regression model. Describe how the regression analysis assisted the authors in reaching a better understanding about their problem and how their study could be improved.

24.15 For a set of time series that is of direct interest to you and for which it would be appropriate to apply regression analysis, explain how you would design a regression model for studying meaningful relationships among the series. Apply the most appropriate regression model to the data set and check that the residual assumptions are satisfied. Explain the advantages and drawbacks of your approach as well as any surprising results that you uncovered.

24.16 In Section 24.3.2, the Spearman partial rank correlation test defined in Section 23.3.6 is employed for checking for trends after removing seasonality. Using an outline similar to the one given in Table 24.3.2, explain how this test can be utilized for taking into account correlation when testing for the presence of a trend in a time series.

24.17 Table 24.3.1 outlines the trend analysis methodology for use with water quality time series measured in rivers. Beyond the techniques referred to in Section 24.3.2, mention other methods that could be employed with this methodology.

24.18 Select a seasonal water quality time series as well as an accompanying riverflow series that are of interest to you. Carry out the methodology of Section 24.2.2 as well as Section 24.3.2 to check for the presence of trends. Clearly explain all of your steps and comment upon your findings.

24.19 In the trend assessment methodology of Section 24.3.2, it is assumed that monthly data is employed in steps 4 to 6 in Table 24.3.1. Carry out the instructions of Problem 24.18 for the case of quarter-yearly data.

24.20 In Steps 4 to 6 in the trend assessment methodology summarized in Table 24.3.1, it is assumed that monthly data are calculated. Execute the instructions of Problem 24.18 for the situation where weekly data are used in these steps.

REFERENCES

APPLICATIONS OF REGRESSION ANALYSIS

Beauchamp, J. J., Downing, D. J., and Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin,* 25(5):961-975.

Cleaveland, M. K. and Durick, D. N. (1992). Iowa climate reconstructed from tree rings, 1640-1982. *Water Resources Research,* 28(10):2607-2615.

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., and Summers, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research,* 28(9):2353-2363.

Duffield, J. W., Neher, C. J., and Brown, T. C. (1992). Recreational benefits of instream flow: Application to Montana's Big Hole and Bitterroot Rivers. *Water Resources Research,* 28(9):2169-2181.

Gunn, J. (1991). Influences of various forcing variables on global energy balance during the period of intense instrumental observation (1958-1987) and their implications for paleoclimate. *Climatic Change,* 19:393-420.

Keppeler, E. T. and Ziemer, R. R. (1990). Logging effects on streamflow: Water yield and summer low flows at Caspar Creek in Northwestern California. *Water Resources Research,* 26(7):1669-1679.

Kite, G. W. and Adamowski, K. (1973). Stochastic analysis of Lake Superior elevations for computation of relative crustal movement. *Journal of Hydrology,* 18:163-175.

Lyman, R. A. (1992). Peak and off-peak residential water demand. *Water Resources Research,* 28(9):2159-2167.

Millard, S. P., Yearsley, J. R., and Lettenmaier, D. P. (1985). Space-time correlation and its effects on methods for detecting aquatic ecological changes. *Canadian Journal of Fisheries and Aquatic Science,* 42:1391-1400.

Porter, P. S. and Ward, R. C. (1991). Estimating central tendency from uncensored trace level measurements. *Water Resources Bulletin,* 27(4):687-700.

Potter, K. W. (1991). Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment. *Water Resources Research,* 27(5):845-855.

See, R. B., Naftz, D. L., and Qualls, C. L. (1992). GIS-assisted regression analysis to identify sources of selenium in streams. *Water Resources Bulletin,* 28(2):315-330.

Simpson, H. J., Cane, M. A., Herczeg, A. L., Zebiak, S. E. and Simpson, J. H. (1993). Annual river discharge in Southeastern Australia related to El Nino - southern oscillation forecasts of sea surface temperature. *Water Resources Research,* 29(11):3671-3680.

Tasker, G. D. (1986). Accounting for unequal record length and cross correlation in regional regression. In *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985, pages 283-290, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

Wong, S. T. (1963). A multivariate statistical model for predicting mean annual flood in New England. *Annals of the Association of American Geographers,* 53(3):298-311.

Wright, K. A., Sendek, K. H., Rice, R. M., and Thomas, R. B. (1990). Logging effects on streamflow: Storm runoff at Caspar Creek in Northwestern California. *Water Resources Research,* 26(7):1657-1667.

BOOKS ON REGRESSION ANALYSIS

Atkinson, A. C. (1985). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications.* Wiley, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Wadsworth and Brooks/Coles, Pacific Grove, California.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis.* Wiley, New York, second edition.

Gallant, A. R. (1987). *Nonlinear Statistical Models.* Wiley, New York.

Helsel, D. R. and Hirsch, R. M. (1992). *Statistical Methods in Water Resources.* Elsevier, Amsterdam.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Massachusetts.

EXPLORATORY DATA ANALYSIS

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Wadsworth, Belmont, California.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts.

FUZZY REGRESSION ANALYSIS

Bardossy, A. (1990). Notes on fuzzy regression. *Fuzzy Sets and Systems,* 37(1):65-75.

Bardossy, A., Bogardi, I., and Duckstein, L. (1990). Fuzzy regression in hydrology. *Water Resources Research,* 26(7):1497-1508.

Bardossy, A., Duckstein, L., and Bogardi, I. (1992). Fuzzy nonlinear regression of dose response relationship. *European Journal of Operational Research.*

Kacrzyk, J. and Federizzi, M., editors (1992). *Fuzzy Regression Analysis.* Physica Verlag, Heidelberg.

Tanaka, H., Uejima, S., and Asai, K. (1982). Linear regression analysis with fuzzy model. *IEEE Transactions on Systems, Man and Cybenetics*, SMC-12:903-907.

POINT OF CHANGE IN A REGRESSION MODEL

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37:149-192.

El-Shaarawi, A. H. and Delorme, L. D. (1982). The change-point problem for a sequence of binomial random variables. In El-Shaarawi, A. H. and Esterby, S. R., editors, *Time Series Methods in Hydrosciences*, pages 68-75. Elsevier, Amsterdam, The Netherlands.

El-Shaarawi, A. H. and Esterby, S. R. (1982). Inference about the point of change in a regression model with a stationary error process. In El-Shaarawi, A. H. and Esterby, S. R., editors, *Time Series Methods in Hydrosciences*, pages 55-67. Elsevier, Amsterdam, The Netherlands.

Esterby, S. R. (1985). A program for estimating the point of change and degree in polynomial regression. Technical Report Scientific Series No. 147, Inland Waters Directorate, National Water Research Institute, Burlington, Ontario, Canada.

Esterby, S. R. and El-Shaarawi, A. H. (1981a). Likelihood inference about point of change in a regression regime. *Journal of Hydrology*, 53:17-30.

Esterby, S. R. and El-Shaarawi, A. H. (1981b). Inference about the point of change in a regression model. *Applied Statistics*, 30(3):277-285.

MacNeill, I. B. (1985). Detecting unknown interventions with application to forecasting hydrological data. *Water Resources Bulletin*, 21(5):785-796.

ROBUST LOCALLY WEIGHTED REGRESSION SMOOTH

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125-127.

Bodo, B. A. (1989). Robust graphical methods for diagnosing trend in irregularly spaced water quality time series. *Environmental Monitoring and Assessment*, 12:407-428.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3-33.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829-836.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Wadsworth, Monterey, California.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 10:1040-1053.

SPURIOUS CORRELATIONS

Benson, M. A. (1965). Spurious correlation in hydraulics and hydrology. *Journal of the Hydraulics Division*, American Society of Civil Engineers (ASCE), 91(HY4):35-42.

Good, I. J. (1959). A classification of fallacious arguments and interpretations. *Technometrics*, 4:125-132.

Good, I. J. (1978). Fallacies, statistical. In Kruskal, W. H. and Tanur, J. M., editors, *International Encyclopedia of Statistics,* Volume 1, pages 337-349. The Free Press, New York.

Huff, D. (1954). *How to Lie with Statistics.* Norton, New York.

Kenney, B. C. (1982). Beware of spurious self-correlations. *Water Resources Research,* 18(4):1041-1048.

Kite, G. (1989). Some statistical observations. *Water Resources Bulletin,* 25(3):483-490.

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisted. *Journal of the Royal Statistical Society,* Series A, 156:379-392.

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society,* London, 60:489-502.

Wong, S. T. (1979). A dimensionally homogeneous and statistically optimal model for predicting mean annual flood. *Journal of Hydrology,* 42:269-279.

Wong, S. T. and DeCoursey, D. G. (1986). More effective development of hydrologic models using dimensional and multivariate analyses. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985, pages 322-338, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

TECHNIQUES USED WITH REGRESSION ANALYSIS

Alley, W. M. (1988). Using exogeneous variables in testing for monotonic trends in hydrologic time series. *Water Resources Research,* 24(11):1955-1961.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society,* Series B, 26:211-252.

Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society,* Series B, 38(2):189-203.

TREND ASSESSMENT USING REGRESSION ANALYSIS

Alley, W. M. (1988). Using exogeneous variables in testing for monotonic trends in hydrologic time series. *Water Resources Research,* 24(11):1955-1961.

Cunningham, R. B. and Morton, R. (1983). A statistical method for the estimation of trend in salinity in the River Murray. *Australian Journal of Soil Research,* 21:123-132.

El-Shaarawi, A. H., Esterby, S. R., and Kuntz, K. W. (1983). A statistical evaluation of trends in the water quality of the Niagara River. *Journal of Great Lakes Research,* 9(2):234-240.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression spline (MARS). *Journal of the American Statistical Association,* 86(416).

Loftis, J. C., Taylor, C. H., Newell, A. D. and Chapman, P. L. (1991). Multivariate trend testing of lake water quality. *Water Resources Bulletin,* 27(3):461-473.

McLeod, A. I., Hipel, K. W., and Bodo, B. A. (1991). Trend analysis methodology for water quality time series. *Environmetrics,* 2(2):169-200.

Reinsel, G. C. and Tiao, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone. *Journal of the American Statistical Association*, 82(397):20-30.

Smith, E. P. and Rose, K. A. (1991). Trend detection in the presence of covariates: Stagewise versus multiple regression. *Environmetrics*, 2(2):153-168.

Stoddard, J. L. (1991). Trends in Catskill stream water quality: Evidence from historical data. *Water Resources Research*, 27(11):2855-2864.

Whitlatch, E. E. and Martin, M. J. (1988). Identification of monthly trends in urban water use. *Water Resources Bulletin*, 24(1):169-174.

WHITENESS TESTS

Ansley, C. F., Kohn, R. and Shirley, T. S. (1992). Computing p-values for the generalized Durbin-Watson and other invariant test statistics. *Journal of Econometrics*, 54:277-300.

Kohn, R., Shively, T. S. and Ansley, C. F. (1993). Computing p-values for the generalized Durbin-Watson statistic and residual autocorrelations in regression. *Applied Statistics*, 42(1):249-269.

Shively, T. S., Ansley, C. F. and Kohn, R. (1990). Fast evaluation of the Durbin-Watson and other invariant test statistics in time series regression. *Journal of the American Statistical Association*, 85:676-685.

Swed, F. S. and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, 14:66-87.

Wallis, K. F. (1972). Testing for fourth-order autocorrelation in quarterly regression equations. *Econometrica*, 40:617-636.

DATA APPENDIX

DATA ACQUISITION

Table 1.6.2 provides a summary of the rich variety of time series models that are presented in the book. To clearly explain how these classes of models can be fitted to real data sets for addressing a range of practical problems, illustrative applications are provided throughout the chapters. At the end of each chapter, exact references are given for the time series utilized in the applications to allow interested readers to obtain the data from the original publication sources. Most of the time series consist of hydrological and other kinds of environmental observations. Nonetheless, as noted in Section 1.6.1 many of the time series models described in the book can be employed by professionals working in fields outside of hydrology and environmental engineering, for application to their particular kinds of time series.

The authors would like to encourage readers to fit models to their own sets of data and use the applications given in the book as a guide. However, some readers may wish to gain confidence in practical time series analysis by fitting models to time series utilized in the applications. Accordingly, some representative time series are listed in this appendix.

To apply the time series models to data a flexible decision support system (DSS) is required. One such system is the MHTS (McLeod-Hipel Time Series) Package referred to in Section 1.7. Included with this package are many of the time series employed in the practical applications in the book.

Most of the datasets referred to in this book are archived in Statlib. This means they are available in electronic form to anyone who has access to e-mail. To obtain these datasets, send the following one-line message to **statlib@lib.stat.cmu.edu :**

send hipel-mcleod from datasets

Statlib is a system for the distribution of software, datasets and general information of interest to statisticians. For further information, one can contact:

Michael M. Meyer
Computing Services and Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
Tel: (412) 268-3108
mikem@stat.cmu.edu

Many environmental and other government agencies throughout the world generously furnish extensive data listings for little or no cost. In Canada and the United States, for instance, one can obtain extensive hydrological time series on CD ROM by contacting, respectively, the agencies given below.

In Canada:

S. Y. Shiau, D. W. Kirk and J. McIlhinney
Water Resources Branch
Environment Canada
Ottawa, Ontario, Canada K1A 0H3

In the United States of America:

J. R. Slack, A. M. Lumb and J. M. Landwehr
Water Resources Division
U. S. Geological Survey
Reston, Virginia, U. S. A. 22092

DATA LISTING

The data listed below consist of four sets of time series to which ARMA, ARIMA, three types of seasonal, and intervention models are fitted. Included with each listing of a time series are an explanation of the specific type of data, the reference in which the data are published, the number of the figure in which the series is plotted in this book, the type of model fitted to the data, and the numbers of the sections where model construction results are presented for the model fitted to the data set. Moreover, the data are listed sequentially from left to right starting with the top line and continuing on lower lines.

Stationary Nonseasonal Time Series

1. Average annual flows of the St. Lawrence River at Ogdensburg, New York in m^3/s from 1860 to 1957.

   Reference: Yevjevich (1963)
   Time Series Plot: Figures 2.3.1 and II.1
   Model Type: Constrained AR(3) model without ϕ_2
   Model Construction

   Identification: Section 5.4.2
   Estimation: Section 6.4.2
   Diagnostic Checks: Section 7.6.2

7788	8040	7733	7528	7528	6962	7699	6853	7051	7897
7331	6342	6710	7392	6540	7447	7133	7133	7331	6908
6567	7249	7106	7644	7160	7869	7617	6826	6962	7419
7331	6485	6853	6853	6117	6028	6171	6458	6458	6396
6424	6458	6853	6908	6737	6826	6908	7419	6819	6540
6199	6485	7303	6826	6260	6792	6826	6997	7106	6396
6628	6485	6144	6260	6062	5892	6396	6737	7222	7447
6171	6171	5892	5326	5183	5435	6062	6171	6117	5946
6028	6062	7024	6997	6853	7276	7276	7303	6826	6683
7644	7788	7331	7194	7249	7303	6826			

2. Annual Wolfer sunspot numbers from 1770 to 1869

Reference: Waldmeier (1961)
Time Series Plot: Figure 5.4.6
Model Construction

Identification: Section 5.4.3
Estimation: Section 6.4.3
Diagnostic Checks: Section 7.6.3

101	82	67	35	31	7	20	93	154	126
85	68	39	23	10	24	83	132	131	118
90	67	60	47	41	21	16	6	4	7
15	34	45	43	48	42	28	10	8	3
0	1	5	12	14	35	46	41	30	24
16	7	4	2	9	17	36	50	64	67
71	48	28	9	13	57	122	138	103	86
65	37	24	11	15	40	62	99	125	96
67	65	54	39	21	7	4	23	55	94
96	77	59	44	47	31	16	7	38	74

3. Average annual temperature data in degrees celcius for the English midlands from 1723 to 1970.

Reference: Manley (1953)
Model Types: AR(2) or MA(2)
Identification Graphs: Sections 2.5.4 and 3.3.2

9.77	9.27	8.66	9.34	9.94	9.52	9.26	10.04	9.85	9.69
10.47	9.80	9.54	10.30	9.92	9.81	9.20	6.84	9.30	8.36
9.81	8.78	8.81	8.61	9.82	8.77	9.44	9.69	8.42	9.19
9.08	8.83	8.54	8.77	8.95	8.95	10.00	9.83	10.00	9.58
8.93	8.72	8.50	8.62	8.69	8.93	8.77	8.51	8.55	9.15
9.24	9.07	10.09	9.01	9.08	9.20	10.40	9.09	10.20	8.01
9.28	7.83	8.54	8.25	9.28	9.21	8.91	9.44	9.27	9.19
9.09	9.89	8.67	9.02	9.00	9.61	7.89	9.23	9.60	8.95
9.05	9.57	8.97	9.80	8.64	8.84	8.93	8.76	9.67	8.20
8.71	7.75	9.06	7.87	8.89	9.84	9.23	8.55	9.51	10.05
8.37	9.31	9.72	10.07	9.46	10.30	8.16	8.69	10.09	9.47
9.49	10.47	9.55	8.86	8.82	8.05	8.68	8.97	8.71	9.22
9.06	8.59	8.26	10.15	9.22	9.42	9.30	9.10	9.14	9.80
8.37	9.31	8.02	9.08	10.07	9.12	9.61	7.89	9.12	9.17
9.67	8.85	9.69	9.65	9.02	10.38	9.62	8.98	9.05	9.75
8.98	9.30	9.43	9.51	9.17	9.24	7.42	9.09	8.56	9.45
9.02	9.83	8.57	8.69	8.27	8.22	8.99	8.73	8.49	8.17
9.97	9.30	8.65	9.33	9.42	10.07	9.69	9.56	9.11	8.83
9.32	9.00	9.13	9.43	8.84	9.36	8.55	9.12	10.05	9.36
9.78	9.88	8.93	9.18	8.51	9.51	8.48	9.57	10.47	8.67

9.08	9.27	9.17	9.72	9.20	9.57	9.01	9.43	8.99	9.38
9.83	9.99	9.72	9.32	9.57	10.18	9.68	9.05	9.09	9.05
10.03	9.57	10.27	9.45	9.57	10.01	10.62	9.41	9.27	9.09
9.84	9.22	9.28	8.83	10.02	9.42	10.48	9.73	9.94	8.59
8.47	9.47	8.95	9.45	9.61	9.30	9.26	9.57		

Nonstationary Nonseasonal Time Series

4. Average annual water use for New York City in litres per capita per day from 1898 to 1968.

Reference: Salas and Yevjevich (1972)
Time Series Plot: Figures II.2 and 4.3.8
Model Type: ARIMA(0,1,0)
Model Identification: Section 4.3.1

402.8	421.3	431.2	426.2	425.5	423.6	435.7	445.2	450.1	450.1
439.1	419.0	417.9	384.2	385.4	374.4	401.3	382.7	403.5	410.0
454.6	448.2	489.5	476.2	473.2	475.1	476.6	502.7	506.5	499.7
495.5	522.8	537.1	509.1	502.7	500.4	508.4	498.9	507.2	505.0
503.8	511.4	467.9	493.6	470.5	503.5	544.3	553.0	551.9	564.4
567.8	562.1	457.3	500.1	522.0	525.4	511.0	533.4	534.1	562.9
557.2	584.1	582.6	590.5	581.1	583.0	567.1	499.3	493.6	533.7
581.1									

Seasonal Time Series

5. Average Monthly flows of the Saugeen River in m^3/s at Walkerton, Ontario, Canada, from January, 1915, until December, 1976.

Reference: Environment Canada (1977)
Time Series Plot: Figure VI.1
Model Types: Deseasonalized and PAR Models with $\lambda = 0$
Model Construction

Deseasonalized Model: Section 13.4.2
PAR model: Section 14.4

16.03	30.30	35.40	41.91	14.70	9.20	7.96	11.95	18.63	21.69	22.57	23.19
69.38	37.10	35.96	125.73	46.16	35.11	12.37	7.84	6.91	9.29	12.18	24.49
16.40	13.22	76.46	82.97	32.85	27.84	57.77	11.81	7.70	13.31	10.73	8.27
7.90	20.95	95.71	67.11	23.98	12.88	6.80	4.84	11.78	12.71	25.82	46.72
28.60	17.73	90.05	53.52	44.17	13.93	9.46	8.13	7.19	10.34	24.18	25.15
14.72	14.72	115.53	63.43	23.02	12.91	19.45	9.46	7.11	12.74	32.56	43.89
40.78	16.03	109.02	50.69	28.09	19.71	13.20	9.12	7.14	11.33	12.06	23.19
13.96	14.72	69.94	92.03	20.27	13.71	12.37	8.58	8.41	7.65	8.86	9.71
8.98	8.21	42.76	103.36	70.51	21.49	11.07	9.29	10.25	7.50	10.48	25.85
23.73	15.91	40.21	89.20	63.15	18.94	14.61	13.51	12.01	9.32	8.13	19.88
10.14	30.02	72.77	41.06	14.67	12.09	10.00	8.27	8.86	18.38	60.31	26.62

21.66	17.56	27.04	128.84	51.54	21.35	11.07	15.18	13.93	25.23	59.75	27.47
19.11	17.13	87.22	34.55	36.25	21.27	13.96	9.06	8.24	9.94	15.83	29.17
29.45	19.82	67.39	111.29	25.34	14.38	31.71	30.02	13.93	40.21	60.60	58.33
64.28	27.30	128.56	137.05	66.54	19.34	16.65	7.84	5.78	6.65	14.81	13.59
46.44	66.83	55.78	85.80	40.78	23.73	12.83	4.81	3.71	5.30	5.21	8.64
7.28	8.83	14.58	70.23	23.45	10.36	7.90	6.97	6.14	9.37	30.02	47.29
64.56	67.96	47.01	83.53	29.45	11.98	19.00	18.32	24.15	20.59	45.31	53.24
38.23	26.05	30.87	93.45	42.76	15.52	8.10	6.46	5.83	7.87	13.22	28.88
33.98	19.40	45.87	105.90	21.07	9.32	5.49	4.11	5.01	4.56	10.70	8.41
20.53	13.03	73.62	21.46	15.29	22.54	8.27	4.62	4.39	4.81	17.02	12.37
8.50	8.47	63.15	62.86	32.28	12.74	6.20	5.32	9.46	12.88	17.73	39.93
56.92	54.93	20.33	70.51	30.30	10.79	8.27	9.26	7.99	12.35	14.58	11.64
13.73	67.39	96.56	41.34	24.24	15.38	6.43	6.31	7.33	5.44	5.61	5.83
19.26	14.16	39.36	107.60	26.25	12.63	12.77	9.71	5.69	10.70	19.43	10.73
9.20	8.61	8.75	118.65	52.67	35.40	12.71	9.37	13.11	18.35	34.55	44.46
34.26	20.67	15.29	99.11	16.82	9.94	5.78	6.43	7.53	29.45	43.32	34.55
22.65	14.58	102.22	67.39	44.17	40.21	10.51	7.62	15.94	21.75	43.89	25.91
32.28	48.70	98.54	117.80	89.48	30.87	18.43	15.66	11.75	9.97	26.53	16.08
22.65	25.63	35.68	92.03	38.51	18.12	8.95	5.97	7.16	5.61	12.09	10.17
10.14	11.84	94.01	50.97	58.62	40.49	38.23	10.70	16.42	54.37	32.28	18.58
52.95	35.11	112.42	27.64	19.11	11.86	7.22	6.37	5.92	6.82	8.66	10.79
23.36	31.71	37.10	208.41	88.91	48.70	23.62	12.49	10.05	10.08	11.27	19.17
10.62	16.03	139.89	59.75	29.17	11.75	10.19	7.19	5.49	8.13	18.49	11.50
44.46	34.83	87.50	58.62	18.77	10.17	8.95	5.69	7.42	9.77	9.09	75.61
71.64	27.38	58.33	132.24	27.13	19.45	15.38	8.83	7.82	9.12	24.10	52.67
59.75	33.70	75.32	143.85	38.79	20.47	19.65	10.22	17.44	35.11	53.24	35.96
63.71	31.71	51.25	116.38	32.85	13.73	9.29	9.60	7.90	6.43	13.03	29.45
25.29	26.08	83.25	42.19	46.16	44.17	40.49	11.13	11.98	11.24	10.22	18.41
14.02	61.73	98.26	112.13	28.60	19.45	8.52	9.29	17.81	101.37	36.81	26.56
33.41	24.95	73.34	107.60	26.36	18.09	7.19	5.38	4.30	7.96	18.32	17.02
11.81	10.53	32.85	126.01	55.22	18.01	13.20	9.85	15.43	11.67	11.81	31.71
21.86	26.93	49.27	44.46	24.35	21.24	31.43	6.65	16.40	18.63	33.13	48.99
21.18	15.66	30.58	49.84	11.92	7.67	5.64	5.52	6.71	6.68	11.16	11.58
13.28	13.71	34.83	133.66	50.40	17.05	10.96	12.94	12.15	19.85	48.99	34.83
34.26	27.33	23.50	147.81	75.04	36.53	13.39	9.00	7.48	8.61	12.97	8.21
6.14	16.71	46.16	45.31	30.30	23.13	16.42	11.38	9.94	8.61	16.11	24.89
15.09	14.53	41.63	65.98	19.34	9.97	6.94	5.95	6.91	11.10	18.80	15.52
9.85	8.35	63.43	60.88	46.44	14.81	10.65	8.72	7.31	6.74	13.59	9.40
22.99	16.48	44.46	42.76	17.95	9.29	7.79	13.14	6.48	6.46	8.86	21.07
17.30	39.36	30.30	110.72	44.17	12.18	9.03	9.94	10.99	24.75	41.63	61.16
32.56	35.40	70.23	47.86	29.45	20.78	7.39	8.10	7.31	9.06	23.19	43.32
36.25	24.24	39.64	97.41	24.10	51.82	33.41	18.55	16.57	33.13	63.15	61.45
25.99	56.07	70.79	52.10	28.60	16.79	10.73	14.98	15.21	16.74	33.70	46.72
35.40	39.36	47.57	119.21	67.11	25.43	16.48	10.42	7.79	12.97	23.62	16.28
11.16	16.48	19.88	109.59	30.30	13.11	20.53	8.86	16.11	23.25	26.36	29.45
15.23	19.40	41.06	122.90	29.45	18.80	13.56	9.71	9.46	8.66	11.47	26.05
22.00	17.27	21.38	125.44	37.10	21.75	17.56	10.51	8.33	16.23	21.04	30.58

59.47	30.58	95.43	50.97	35.11	24.13	11.47	9.06	6.57	8.55	22.14	23.19
31.15	24.21	75.32	83.82	61.73	19.00	12.69	9.57	7.99	11.44	23.70	16.11
30.30	26.11	50.97	124.59	39.64	16.48	11.24	12.20	19.06	12.80	15.86	39.08
15.60	40.21	150.36	60.88	35.68	16.91	23.64	11.24	16.20	19.85	29.45	19.00

6. Average monthly water consumption in millions of litres per day from 1966 to 1988 for the city of London, Ontario, Canada.

Reference: Public Utilities Commission (1989) of London
Time Series Plot: Figure VI.2
Model Type: SARIMA(1,0,1)×(0,1,1) with $\lambda = -0.75$
Model Construction: Section 12.4.2

76.83	77.74	80.47	79.56	82.28	100.92	113.20	90.92	86.83	82.74
83.65	80.92	83.19	83.65	83.65	83.65	86.83	100.47	91.38	101.38
95.92	88.19	88.19	80.47	80.92	79.56	80.92	88.19	91.83	96.38
97.29	102.29	99.10	92.74	87.29	85.47	91.38	92.74	89.56	88.65
93.20	99.56	109.11	124.56	115.47	96.38	92.29	86.83	87.29	85.92
85.92	88.65	91.83	112.29	101.83	125.02	102.74	95.01	91.83	86.38
87.29	88.19	89.10	89.10	103.65	127.75	125.47	125.47	109.11	100.01
95.01	85.01	86.83	86.83	86.83	86.83	100.47	111.38	105.47	102.74
105.01	96.38	94.10	86.83	92.74	93.20	95.47	96.38	99.56	120.47
123.20	114.11	120.93	102.74	101.83	95.47	100.01	100.01	98.20	100.01
103.65	114.56	134.11	131.84	113.65	107.29	102.29	94.56	97.29	98.20
95.47	100.47	116.38	117.29	140.93	120.02	111.38	108.65	105.92	99.10
101.83	102.74	102.74	105.47	108.65	139.57	110.47	118.65	120.02	109.11
108.20	101.38	106.38	108.65	107.74	105.92	129.56	139.11	125.93	123.65
118.65	110.47	110.02	100.47	104.1	106.6	105.5	107.5	117.9	136.3
156.8	135.8	130	117.5	115.8	105.5	111.6	113.2	113.1	112.5
120	147.6	149.9	131.2	134.6	122.2	117.7	106.8	111.5	111.3
109.5	112.1	127	135.9	150.4	135.6	134.9	124.1	120.8	112.8
117.4	118.6	119.2	119.7	128.6	142.8	170	145.9	140.1	128.7
123.4	114.6	120.2	122	121.3	123.2	141.1	129.7	152.4	141.9
137	129	124.6	117.3	122.7	121	122	122	126.3	158.1
164.9	143.3	151.4	136.8	133.1	124.8	132.6	130.2	129.6	129.7
133.7	148.3	155.1	157.2	147.2	142.7	135.9	123.8	132.3	132.7
130.7	129.9	145.5	156.6	161.7	156	146.1	136.8	132.5	129.5
129.5	134.7	136.6	138.4	149.6	159.5	171.4	162.1	163.1	152.4
145.5	133.9	136.6	139.4	141.2	144.9	181.4	187	211.4	178.1
168	154.4	150.4	139.4	144.7	143	148.3	152.7	173.3	226.3
218.2	184.6	174.9	161.4	161.4	145.8				

7. Average monthly concentrations of atmospheric CO_2 measured in molefractions in ppm at the Mauna Loa Observatory in Hawaii from January, 1965, to December, 1980.

   Reference: Keeling et al. (1982) and Bacastow and Keeling (1981)
   Time Series Plot: Figure VI.3
   Model Type: SARIMA$(0,1,1)\times(0,1,1)_{12}$
   Model Construction: Section 12.4.3

319.32	320.36	320.82	322.06	322.17	321.95	321.20	318.81	317.82	317.37
318.93	319.09	319.94	320.98	321.81	323.03	323.36	323.11	321.65	319.64
317.86	317.25	319.06	320.26	321.65	321.81	322.36	323.67	324.17	323.39
321.93	320.29	318.58	318.60	319.98	321.25	321.88	322.47	323.17	324.23
324.88	324.75	323.47	321.34	319.56	319.45	320.45	321.92	323.40	324.21
325.33	326.31	327.01	326.24	325.37	323.12	321.85	321.31	322.31	323.72
324.60	325.57	326.55	327.80	327.80	327.54	326.28	324.63	323.12	323.11
323.99	325.09	326.12	326.61	327.16	327.92	329.14	328.80	327.52	325.62
323.61	323.80	325.10	326.25	326.93	327.83	327.95	329.91	330.22	329.25
328.11	326.39	324.97	325.32	326.54	327.71	328.73	329.69	330.47	331.69
332.65	332.24	331.03	329.36	327.60	327.29	328.28	328.79	329.45	330.89
331.63	332.85	333.28	332.47	331.34	329.53	327.57	327.57	328.53	329.69
330.45	330.97	331.64	332.87	333.61	333.55	331.90	330.05	328.58	328.31
329.41	330.63	331.63	332.46	333.36	334.45	334.82	334.32	333.05	330.87
329.24	328.87	330.18	331.50	332.81	333.23	334.55	335.82	336.44	335.99
334.65	332.41	331.32	330.73	332.05	333.53	334.66	335.07	336.33	337.39
337.65	337.57	336.25	334.39	332.44	332.25	333.59	334.76	335.89	336.44
337.63	338.54	339.06	338.95	337.41	335.71	333.68	333.69	335.05	336.53
337.81	338.16	339.88	340.57	341.19	340.87	339.25	337.19	335.49	336.63
337.74	338.36								

Time Series Containing an Intervention

8. Average annual flows in m^3/s of the Nile River at Aswan, Egypt. Average yearly values are calculated for the water year from October 1 to September 30 for each year from October 1, 1870, to September 30, 1945. From 1903 onwards there was a drop in the mean level of the Nile flows because of the construction of the Aswan Dam.

   Reference: Hurst et al. (1946)
   Time Series Plot: Figure 19.2.1
   Model Type: Intervention model having a step intervention and an AR(1) noise term
   Model Construction: Section 19.2.4

3958.043	3369.694	3485.242	3437.691	3702.352	3817.610
2875.578	3054.686	4724.150	3834.007	3076.773	2965.759
3461.708	3141.010	3371.237	2988.425	3607.541	2946.083
2709.200	3294.848	3556.615	3653.934	3846.064	3713.637
4252.313	3657.503	3639.370	3197.722	3112.749	2353.684
2843.652	2194.926	2689.428	2950.906	2247.877	2628.279
2491.126	2792.630	3321.469	3058.062	2889.853	2495.273

1648.823	1981.963	2411.072	3035.203	3556.133	3261.959
2377.893	2394.964	2499.999	2610.242	2743.633	2744.116
2338.637	2494.984	2474.440	2446.373	2963.059	2732.252
2205.150	2681.808	2580.535	2954.378	3025.944	2902.777
2642.457	2860.242	2665.412	2306.905	1848.090	2569.540
2503.954	2438.753	2211.130			

9. Average monthly phosphorous data in mg/l from January, 1972, until December, 1977, for measurements taken by the Ontario Ministry of the Environment downstream from the Guelph sewage treatment plant located in the Grand River basin, Ontario, Canada. In February, 1974, a pollution abatement procedure was brought into effect by implementing conventional phosphorous treatment at the Guelph station. The man-induced intervention of phosphorous removal decreased the mean level of the series after the intervention date. Values that are underlined indicate where there are missing data points. The value written above a line is the monthly average across all of the years.

Source: Ontario Ministry of the Environment, Toronto, Ontario, Canada
Time Series Plot: Figures 1.1.1 and 19.1.1
Model Type: Intervention model having a step intervention component, four missing value terms and a SARMA(0,5)×$(0,1)_{12}$ noise term fitted to the logarithmic data.

.4700	.5100	.3500	.1900	.3300	.1524	.3650	.6500	.8250	1.0000	.3850	.9000
.2950	.1400	.2200	.2000	.1400	.4000	.2144	.4950	1.1000	.5900	.2700	.3000
.3064	.0650	.2400	.0580	.0790	.0650	.1200	.0910	.0580	.1200	.1200	.1100
.4600	.1500	.0860	.0280	.1342	.1100	.3600	.1800	.0650	.1300	.1200	.1900
.1500	.1070	.0470	.0550	.0800	.0710	.1210	.1080	.1690	.0660	.0790	.1040
.1570	.1400	.0700	.0560	.0420	.1160	.1060	.0940	.0970	.0500	.0790	.1140

REFERENCES

Bacastow, R. B. and Keeling, C. D. (1981). Atomospheric carbon dioxide concentration and the observed airborne fraction. In Bolin, B., editor, *Carbon Cycle Modelling*, pages 103- 112. John Wiley, Chichester.

Environment Canada (1977). *Historical streamflow summary, Ontario.* Thechnical report, Inland Water Directorate, Water Resources Branch, Ottawa, Canada.

Hurst, H. E., Black, R. P. and Simaika, Y. M. (1946). *The Nile Basin, Volume VII, The future conservation of the Nile.* Ministry of Public Works, Physical Department Paper No. 51, S.O.P. Press, Cairo, Egypt.

Keeling, C. D., Bacastow, R. B. and Whorf, T. P. (1982). Measurements of the concentration of carbon dioxide at Mauna Loa Observatory, Hawaii. In Clark, W. C., editor, *Carbon Dioxide Review 1982*, pages 377- 385. Clarendon Press, Oxford.

Manley, G. (1953). The mean temperatures of central England (1698 – 1952). *Quarterly Journal of the Royal Meteorological Society*, 79: 242- 261.

Public Utilities Commission (1989). *Water usage data for the city of London, Ontario.* Technical report, Public Utilities Commission, P.O.Box 2700, London, Ontario.

Salas, J. D. and Yevjevich, V. M. (1972). *Stochastic structure of water use time series.* Hydrology Paper No. 52, Colorado State University, Fort Collins, Colorado.

Waldmeier M. (1961). *The Sunspot Activity in the Year 1610–1960.* Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). *Fluctuation of wet and dry years, 1, research data assembly and mathematical models.* Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

AUTHOR INDEX

This Page Intentionally Left Blank

SUBJECT INDEX

This Page Intentionally Left Blank