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**.

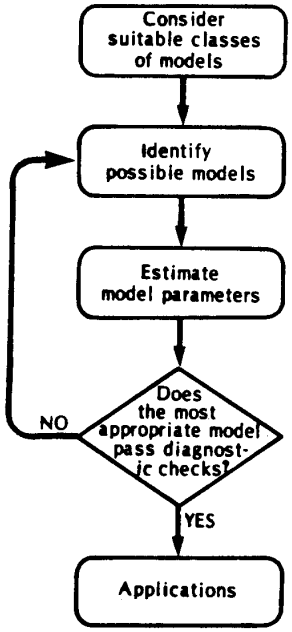

Figure III.1. Model construction.

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

$$var(ri_k) \approx \frac{1}{n}\left\{1 + 2\sum_{j=1}^{p} ri_j^2\right\}, \quad k > p \qquad [5.3.3]$$

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.

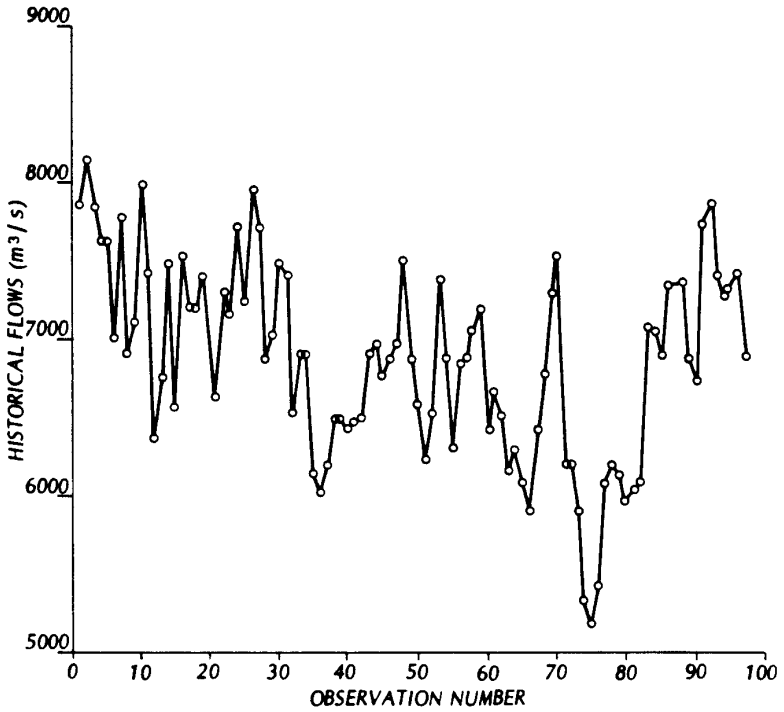

Figure 5.4.1. Annual flows of the St. Lawrence River at Ogdensburg, New York from 1860 to 1957.

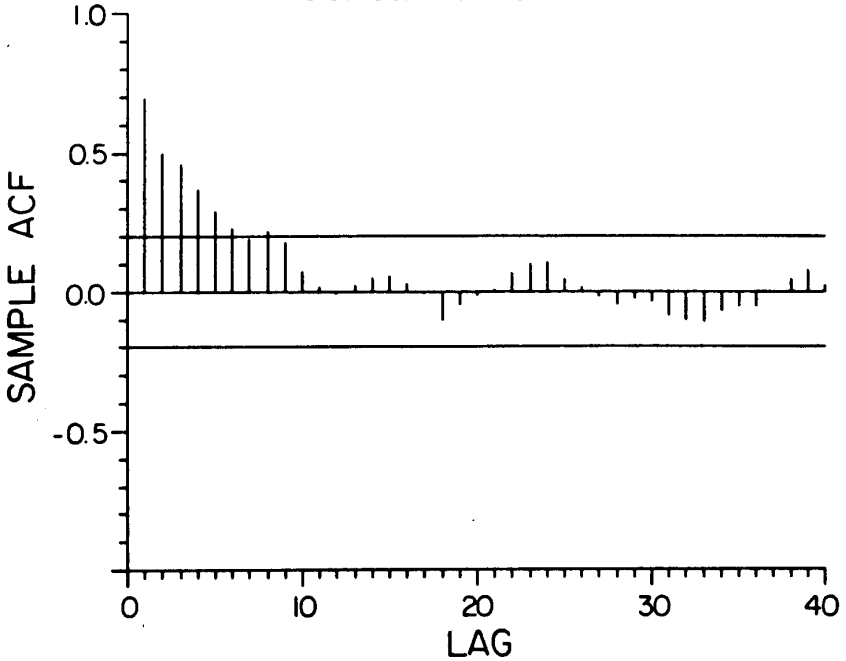

Figure 5.4.2. Sample ACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

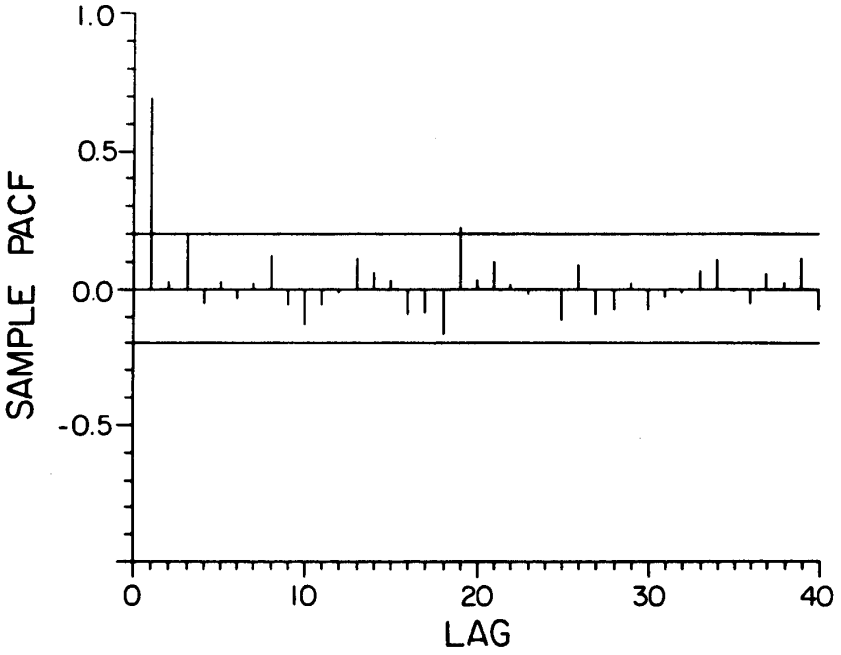

Figure 5.4.3. Sample PACF and 95% confidence limits for the average yearly flows of the St. Lawrence River at Ogdensburg, New York.

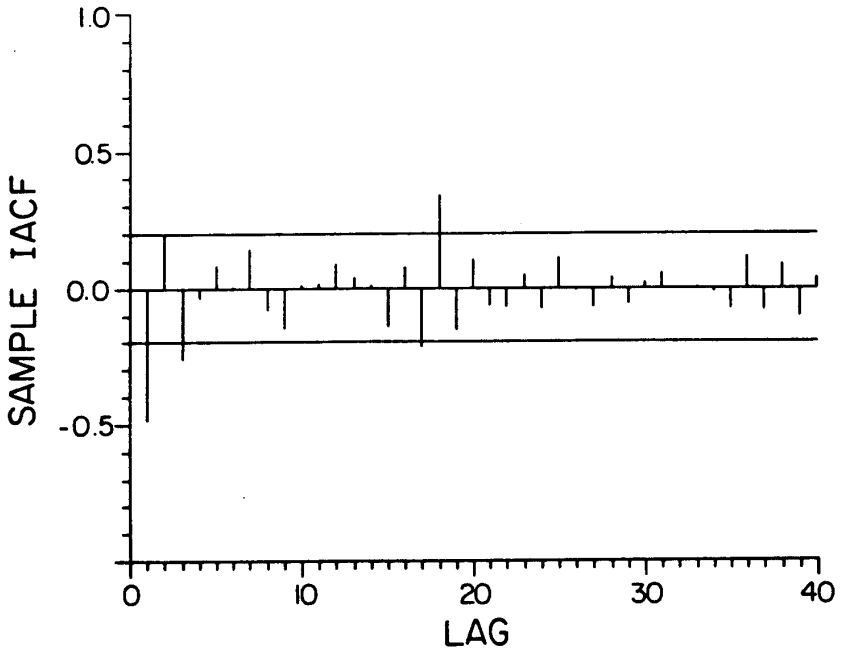

Figure 5.4.4. Sample IACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

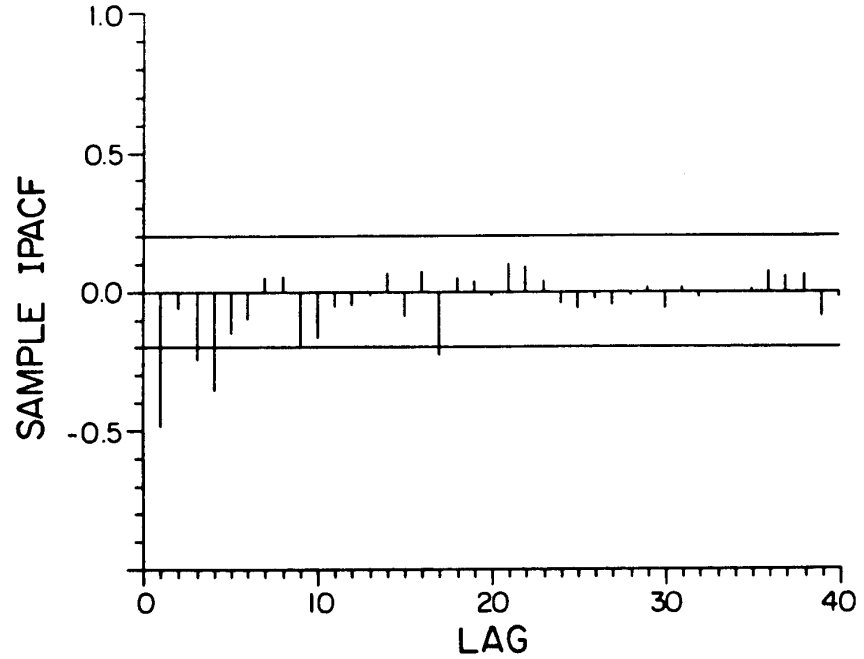

Figure 5.4.5. Sample IPACF and 95% confidence limits for the average annual flows of the St. Lawrence River at Ogdensburg, New York.

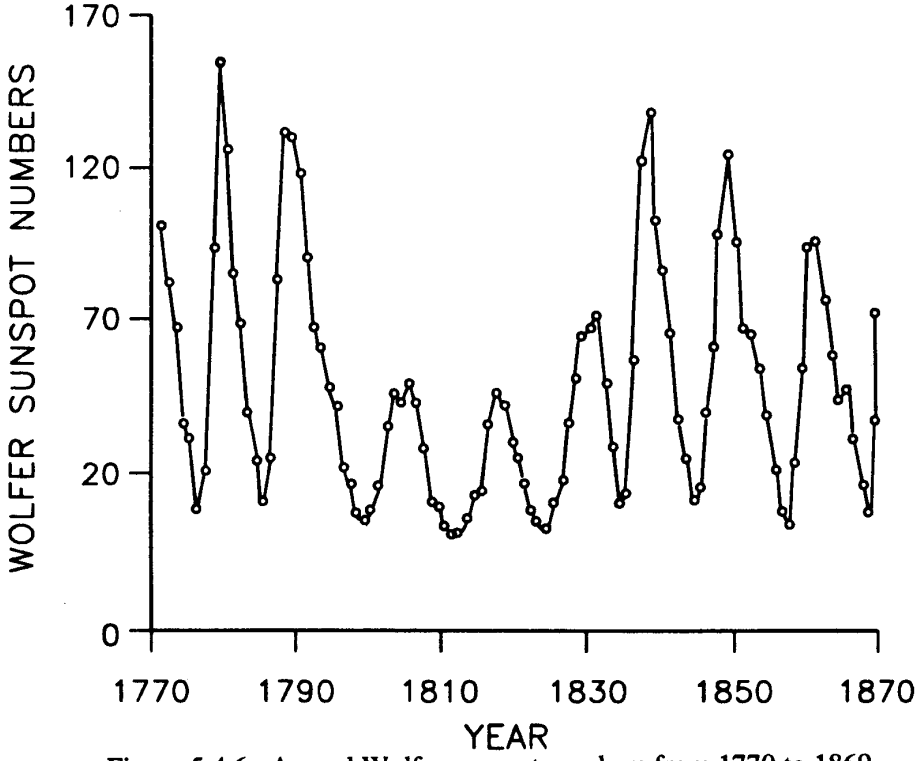

Figure 5.4.6. Annual Wolfer sunspot numbers from 1770 to 1869.

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.

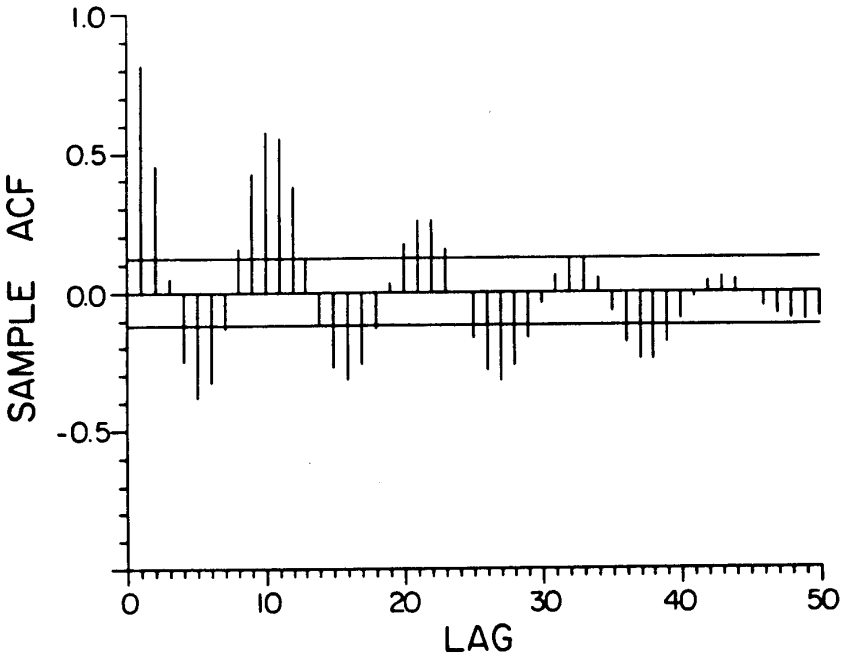

Figure 5.4.7. Sample ACF and 95% confidence limits for the yearly sunspot numbers.

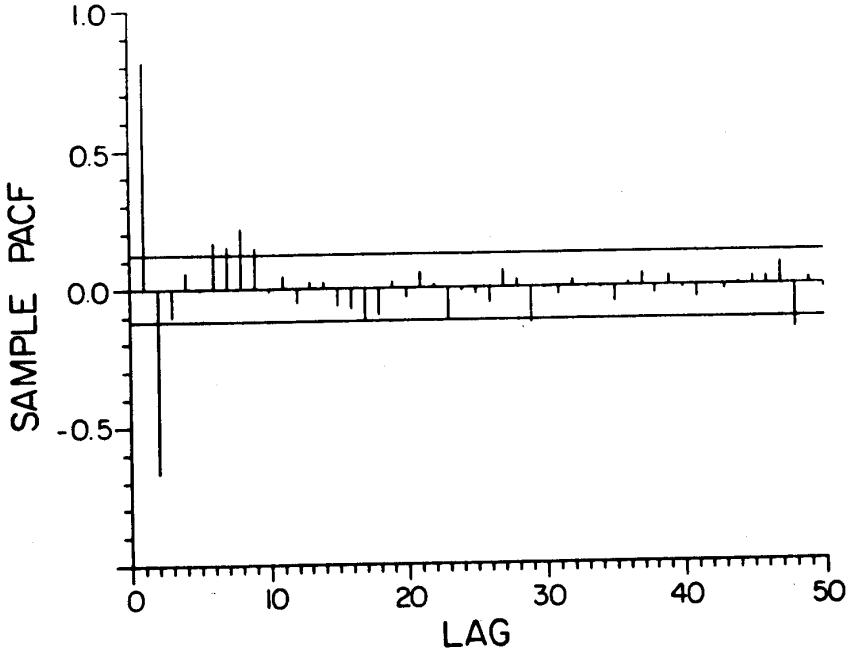

Figure 5.4.8. Sample PACF and 95% confidence limits for the annual sunspot numbers.

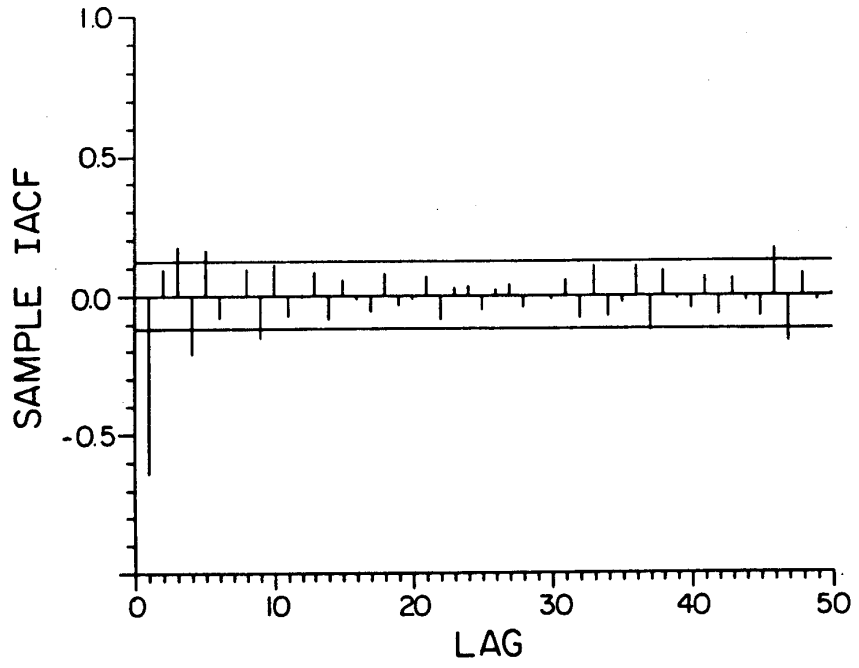

Figure 5.4.9. Sample IACF and 95% confidence limits for the yearly sunspot numbers.

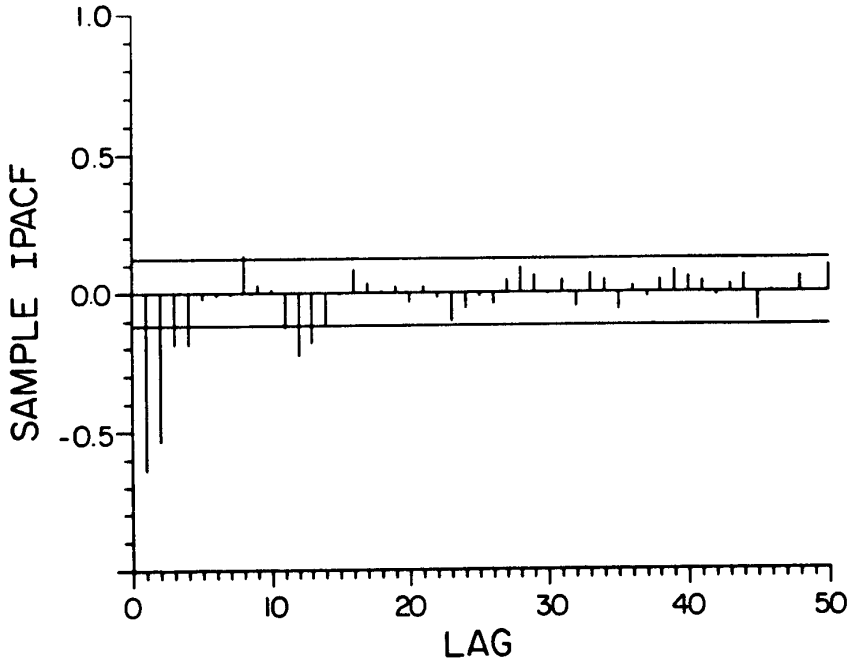

Figure 5.4.10. Sample IPACF and 95% confidence limits for the annual sunspot numbers.

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.