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

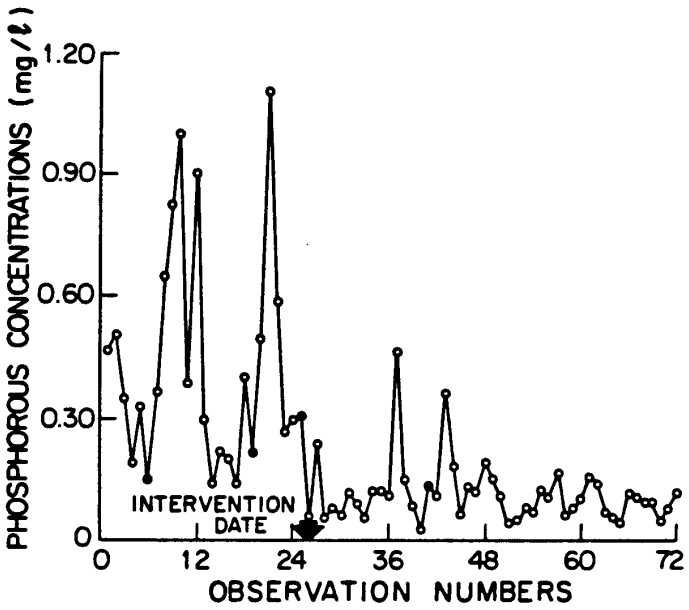


Figure 19.1.1. Monthly phosphorous data (mg/*l*) on the Speed River near Guelph, Ontario, Canada.

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

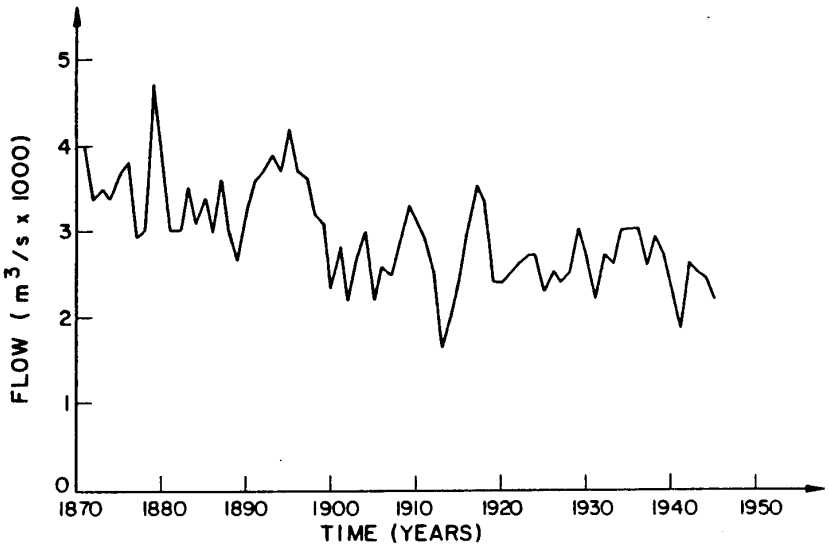


Figure 19.2.1. Average annual flows of the Nile River at Aswan.

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, $\mathbf{k}$, 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.

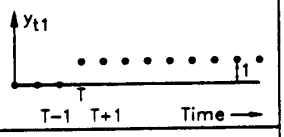
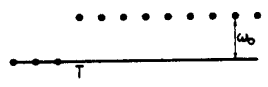
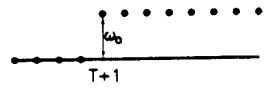
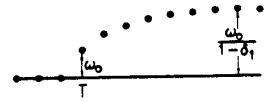
Figure	$\frac{\omega(B)}{\delta(B)} B^b S_t^{(T)}$	Graph of Dynamic Response to a Step Input
a	$S_t^{(T)}$	
b	$\omega_0 S_t^{(T)}$	
c	$\omega_0 B S_t^{(T)}$	
d	$\frac{\omega_0}{1-\delta_1 B} S_t^{(T)}$	
e	$\frac{\omega_0 B}{1-\delta_1 B} S_t^{(T)}$	

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

$$\begin{aligned} \nu(B) &= \frac{\omega_0 B}{1-\delta_1 B} = \omega_0 B(1-\delta_1 B)^{-1} \\ &= \omega_0 B(1+\delta_1 B+\delta_1^2 B^2+\delta_1^3 B^3+\cdots) \\ &= \omega_0(B+\delta_1 B^2+\delta_1^2 B^3+\delta_1^3 B^4+\cdots) \end{aligned} \qquad [19.2.4]$$

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, $\nu_0,\nu_1,\nu_2,\nu_3,\ldots,$, can be obtained directly from [19.2.4] by comparing coefficients of B^k , k=0,1,2, . . . , in $\nu(B) = \nu_0 + \nu_1 B + \nu_2 B^2 + \nu_3 B^3 + \cdots$, to those in [19.2.4]. Consequently, $\nu_0 = 0$, $\nu_1 = \omega_0$, $\nu_2 = \omega_0\delta_1$, $\nu_3 = \omega_0\delta_1^2$, and in general $\nu_k = \omega_0\delta_1^{k-1}$. 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 $(1-B)^{-1}$.

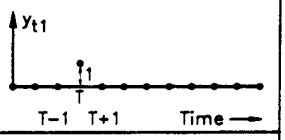
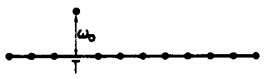
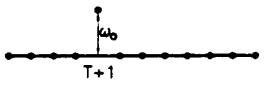
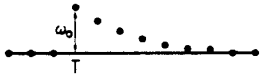
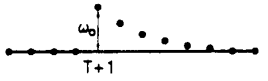
Figure	$\frac{\omega(B)}{\delta(B)} B^b P_t^{(T)}$	Graph of Dynamic Response to a Pulse Input
a	$P_t^{(T)}$	
b	$\omega_0 P_t^{(T)}$	
c	$\omega_0 B P_t^{(T)}$	
d	$\frac{\omega_0}{1-\delta_1 B} P_t^{(T)}$	
e	$\frac{\omega_0 B}{1-\delta_1 B} P_t^{(T)}$	

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

$$f(\mathbf{k},\xi,t) = f(\delta,\omega,\mathbf{b},\xi,t)$$
$$= \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} \qquad [19.2.6]$$

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

$$\nu_i(B) = \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}$$
$$= \frac{(\omega_{0i} - \omega_{1i}B - \omega_{2i}B^2 - \cdots - \omega_{m_i i}B^{m_i})B^{b_i}}{(1 - \delta_{1i}B - \delta_{2i}B^2 - \cdots - \delta_{r_i}B^{r_i})}$$

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 $\nu_i(B)$ 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 ν_{ji}, j=0,1,2, . . . , in the operator

$$\nu_i(B) = \nu_{0i} + \nu_{1i}B + \nu_{2i}B^2 + \cdots$$
$$= \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}$$

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

$$\begin{aligned} y_t - \mu_y &= \nu(B)\xi_{st} + N_t \\ &= \frac{\omega(B)}{\delta(B)}B^b\xi_{st} + \frac{\theta(B)}{\phi(B)}a_t \end{aligned} \qquad [19.2.8]$$

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

$$\begin{aligned} y_t - \mu_y &= \sum_{i=1}^{I_1}\nu_i(B)\xi_{sti} + N_t \\ &= \sum_{i=1}^{I_1}\frac{\omega_i(B)}{\delta_i(B)}B^{b_i}\xi_{sti} + \frac{\theta(B)}{\phi(B)}a_t \end{aligned} \qquad [19.2.9]$$

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

$$Y_t = e^{\mu_y}e^0e^{N_t} = e^{\mu_y}e^{N_t}$$

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.

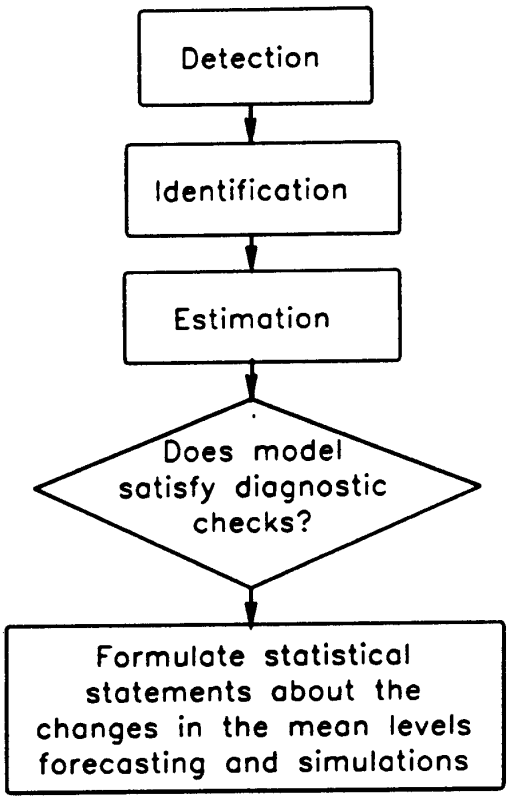


Figure 19.2.4. Constructing an intervention model.

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.

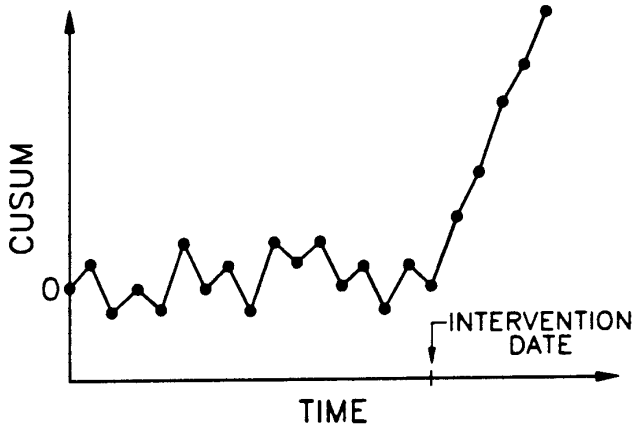


Figure 19.2.5. Cusum chart for a step increase in mean.

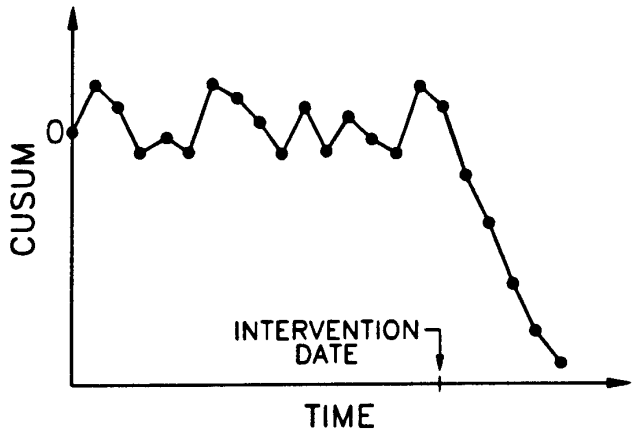


Figure 19.2.6. Cusum chart for a step decrease in mean.

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.

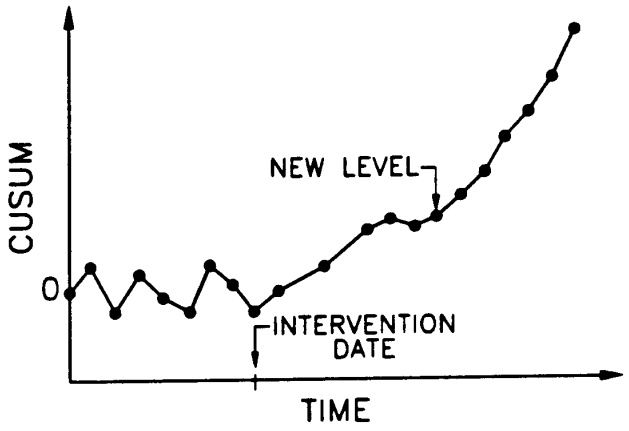


Figure 19.2.7. Cusum chart for a gradual increase in mean to a new level.

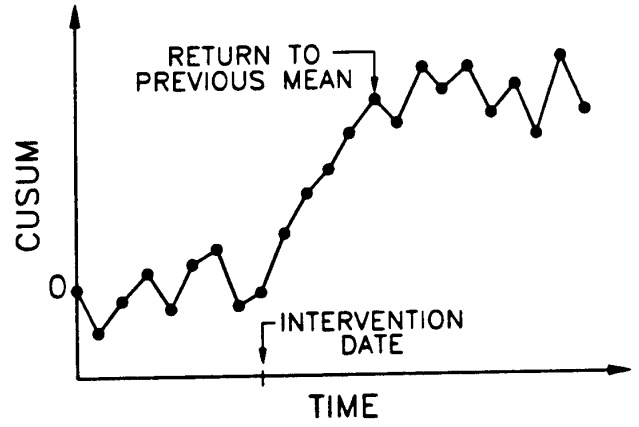


Figure 19.2.8. Cusum chart for a step increase in the mean followed by a step decrease to the previous average.

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

$$y_t - \mu_y = \nu(B)S_t^{(T)} + N_t$$

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.

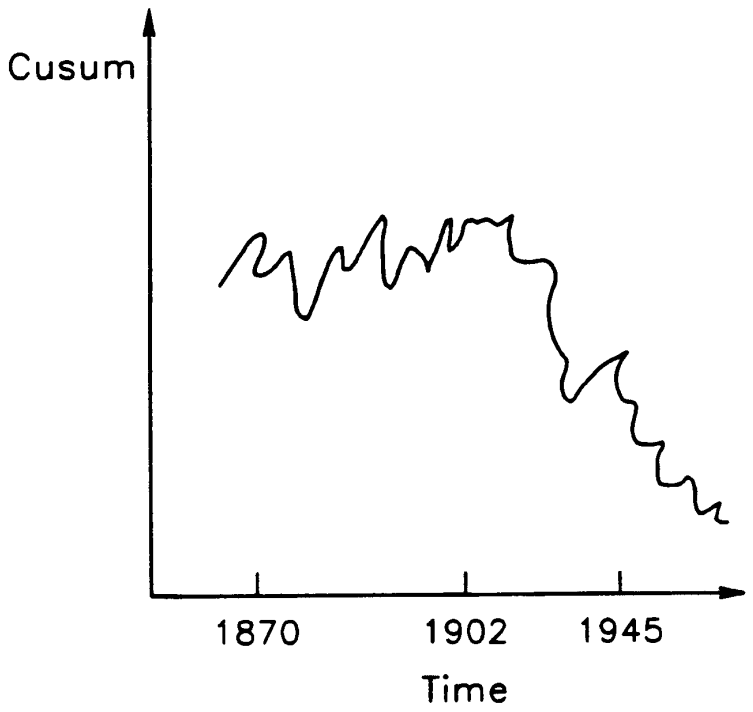


Figure 19.2.9. Cusum for the average annual flows of the Nile River.

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 $\nu(B)$		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

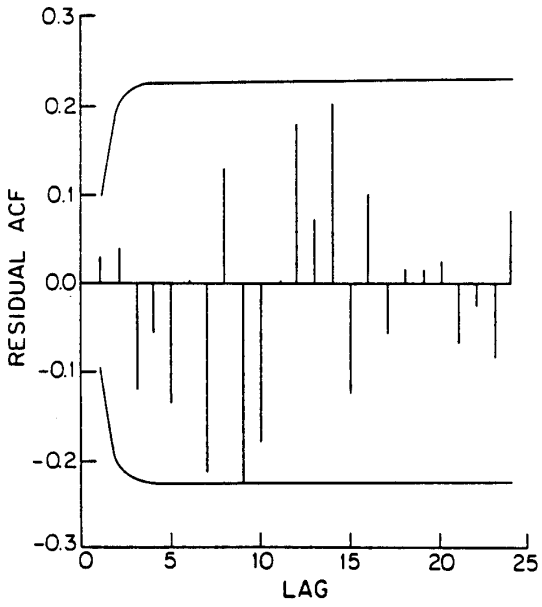


Figure 19.2.10. Residual ACF for the Nile River intervention model.

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;

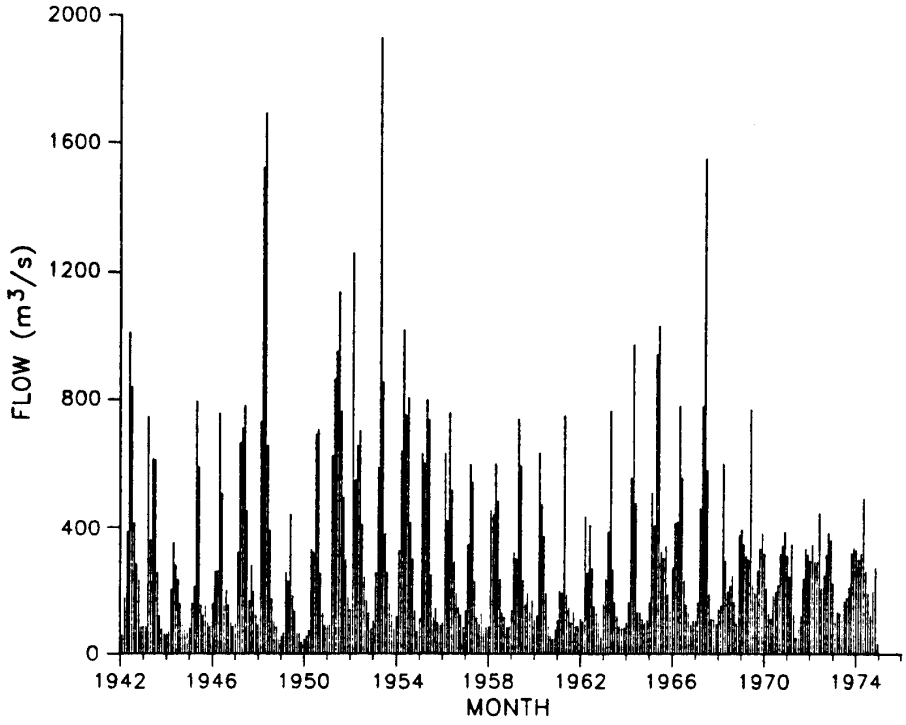


Figure 19.2.11. Average monthly flows of the S. Sask. River.

$$\xi_{ti} = \begin{cases} 1, & t = ith\ month \text{ for } all\ years\ after\ 1968 \\ 0, & otherwise \end{cases}$$

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

$$y_t - \bar{y} - \left[\sum_{i=1}^{12} \omega_{0i}\xi_{it}\right] = N_t$$

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:

$$y_t - \bar{y} = x_t + \sum_{i=1}^{12} \omega_{0i}\xi_{xi} + \frac{a_t}{l - \phi_1 B} \qquad [19.2.24]$$

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

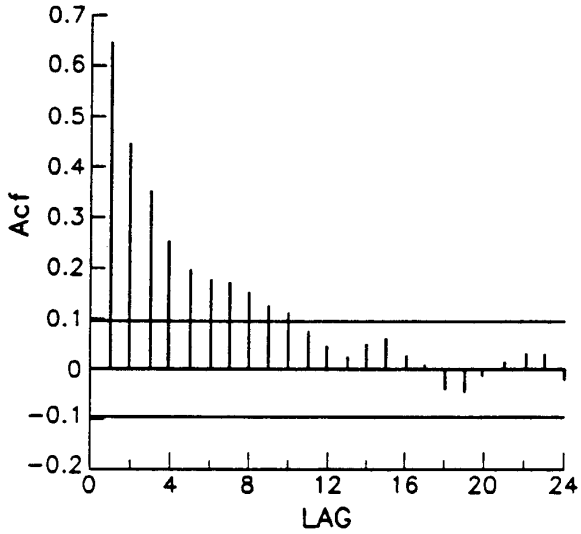


Figure 19.2.12. ACF and 95% confidence limits for the S. Sask. River residuals.

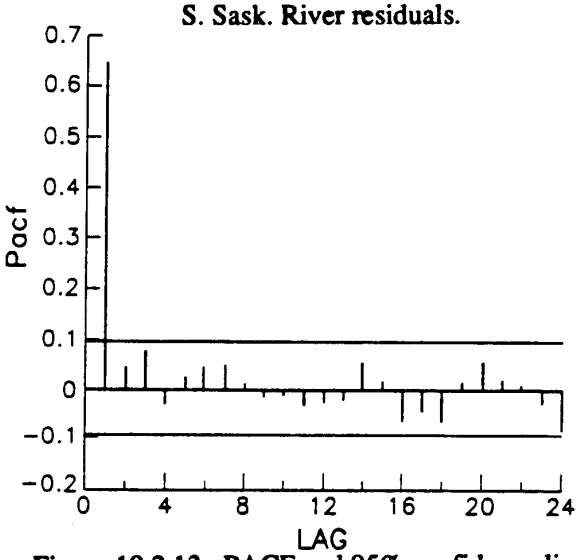


Figure 19.2.13. PACF and 95% confidence limits for the S. Sask. River residuals.

$$y_t = e^{\bar{y}} e^{x_t} e^{N_t} \exp\left[\sum_{i=1}^{12} \omega_{0i} \xi_{it}\right]$$

$$= c_1 e^{x_t} e^{N_t} \exp\left[\sum_{i=1}^{12} \omega_{oi} \xi_{it}\right]$$

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

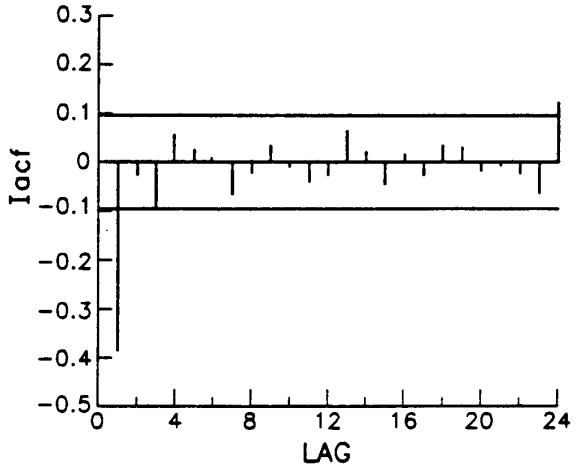


Figure 19.2.14. IACF and 95% confidence limits for the S. Sask. River residuals.

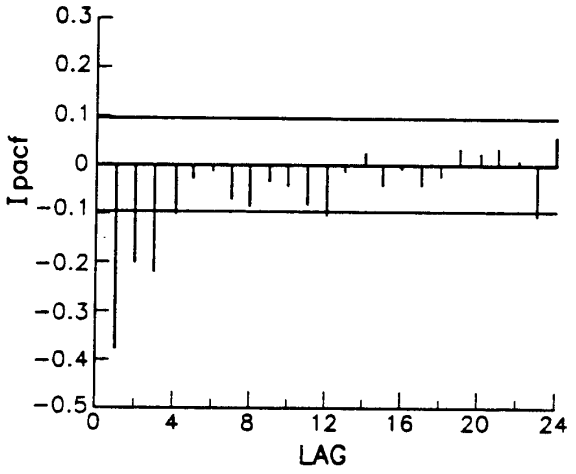


Figure 19.2.15. IPACF and 95% confidence for the S. Sask. River residuals.

$$E[Y_t]_{before} = c_1 c_2$$

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 $\tilde{y}_{t_1}$ where, for instance, $\tilde{y}_{t_1}$ 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

$$\nu_j(B)\xi_{tj} = \omega_{0j}\xi_{tj}$$

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

$$a_t = (y_t - \bar{y}) - \sum_{j=1}^{l_2} \omega_{0j}\xi_{tj}$$

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

$$(1 - \phi_1 B - \phi_3 B^3)(Y_t - \mu_y) = a_t$$

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

$$-\hat{\omega}_{01} = \bar{Y} + \frac{1}{1 - \hat{\phi}_1 B - \hat{\phi}_3 B^3} \hat{a}_{33}$$

and

$$-\hat{\omega}_{02} = \bar{Y} + \frac{1}{1 - \hat{\phi}_1 B - \hat{\phi}_3 B^3} \hat{a}_{34}$$

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,

$$-\hat{\omega}_{01} = \bar{y} + \frac{1}{1 - \hat{\phi}_1 B - \hat{\phi}_3 B^3} \hat{a}_{25}$$

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

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.4.2]$$

The first summation term on the right hand side accounts for the I_1 external interventions modelled in Section 19.2.2 where $\nu_i(B)$ 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

$$\nu_j(B)\xi_{ij} = \omega_{oj}\xi_{ij}$$

where ω_{0j} is the only required transfer function parameter and ξ_{ij} 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

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{i=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + a_t$$

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	-0.1467	0.2870	0.4657	0.3303	-0.3460	-1.3720
(0.1113)	(0.1014)	(0.0971)	(0.1027)	(0.1128)	(0.1138)	(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

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + N_t \qquad [19.5.2]$$

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, $\nu_i(B)$, which is the same as the one defined in [17.5.2] for TFN models, is given as

$$\nu_i(B) = \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}$$

$$= \frac{\omega_{0i} - \omega_{1i}B - \omega_{2i}B^2 - \cdots - \omega_{m_i i}B^{m_i}}{1 - \delta_{1i}B - \delta_{2i}B^2 - \cdots - \delta_{r_i i}B^{r_i}}B^{b_i}$$

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 $\nu_i(B)$ 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 $\nu_i(B)\xi_{ti}$, 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

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + N_t \qquad [19.5.6]$$

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

$$y_t - \mu_y = \sum_{k=1}^{I_3} \nu_k(B)(x_{tk} - \mu_{xk}) + N_t \qquad [19.5.7]$$

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.

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + \sum_{k=I_1+I_2+1}^{I_1+I_2+I_3} \nu_k(B)(x_{tk} - \mu_{xk}) + N_t \qquad [19.5.8]$$

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

$$\nu_j(B)\xi_{tj} = \omega_{0j}\xi_{tj}$$

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 $\nu_k(B)$. 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]

$$N_t = (y_t - \mu_y) - \left(\sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} \right)$$

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, $\nu_k(B)$ 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

$$\nu_k(B) = \omega_{0k}$$

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

$$y_t - \mu_y = \sum_{i=1}^{I_1} \nu_i(B)\xi_{ti} + \sum_{j=I_1+1}^{I_1+I_2} \omega_{0j}\xi_{tj} + \sum_{k=I_1+I_2+1}^{I_1+I_2+I_3} \nu_k(B)(x_{tk} - \mu_{tk}) + a_t$$

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

$$\textit{Pipers Hole flows} = \textit{dynamic component} + \textit{noise}$$

where

$$\textit{dynamic component} = \textit{fire intervention} + \textit{Bay du Nord flows}$$

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

$$E[Y_t]_{\text{before}} = c'_1 c'_2 \qquad [19.5.15]$$

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

$$E[Y_t]_{\text{after}} = c'_1 c'_2 e^{\omega_{01}\delta_{11}^{(\text{date}-1962)}} \qquad [19.5.16]$$

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 $-5\,^{o}C$. 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. 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

$$y_{r,m} - \mu_m = \omega_{0m}\xi_{tm} + \frac{a_{r,m}}{1 - \phi_{1,m}B - \phi_{2,m}B^2} \qquad [19.6.1]$$

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

$$\textit{response variable} = \textit{dynamic component} + \textit{noise}$$

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

   $$\textit{dynamic component} = \textit{interventions}$$

2. If there are missing data in a series, the procedure of Section 19.3 can be used where

   $$\textit{dynamic component} = \textit{missing data}$$

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

   $$\textit{dynamic component} = \textit{interventions} + \textit{missing data}$$

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 yearly 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.