PART X

HANDLING MESSY ENVIRONMENTAL DATA

In an **environmental impact assessment** study, an analyst may be requested to detect and model trends in a data set provided by the client. Unfortunately, as discussed in detail in Section 19.3.1, there are many reasons as to why the quality of environmental data, such as a set of water quality time series, is often not very high. One of the major problems with environmental series is there are often missing data points among which there may be long periods of time for which no observations were taken. In addition, there may be one or more external interventions which affect the stochastic manner in which a series behaves. In other words, **environmental data are often quite messy.**

The major objective of Part X is to explain how an optimal amount of information from messy environmental series can be detected and modelled. To accomplish this, the data analysis methodology of Tukey (1977) can be followed. As initially mentioned in Section 1.2.4, the two main steps in an overall **data analysis** study are:

1. **Exploratory Data Analysis** (Section 22.3 as well as Sections 5.3.2, 19.2.3 and 24.2.2);
2. **Confirmatory Data Analysis** (Section 22.4 as well as Chapter 3 to 21 and Sections 23.3 and 24.2.3).

In Section 22.3, a range of useful exploratory data analysis tools are suggested for discovering important patterns and statistical characteristics such as trends, caused by external interventions. To demonstrate the insights that can be gained by employing exploratory data analysis tools, they are applied to water quality series in Sections 22.3 23.5, 24.2.2 and 24.3.2 within Part X, as well as many other locations in the book.

To rigorously characterize trends and other desired statistical traits which may be known in advance or else detected using exploratory data analysis studies (see Section 19.2.3), formal statistical techniques can be employed at the confirmatory data analysis stage. In Part X, the following three types of confirmatory data analysis methods are described and used in environmental applications:

1. **Intervention Analysis** (Section 22.4 and also Part VIII),
2. **Nonparametric Tests** (Chapter 23),
3. **Regression Analysis** (Section 24.2.3).

As explained in detail in Chapter 19 and exemplified by applications throughout Chapter 19 and in Section 22.4, **intervention analysis** can be used to ascertain the magnitudes of changes in the mean levels of a series due to one or more external interventions. To determine the most appropriate intervention model to fit to a given data set at the confirmatory data analysis stage, one can follow the identification, estimation and diagnostic check stages of model construction. However, in order to be able to fit an intervention model to the time series, a sequence of data points evenly spaced in time must be available. If there are missing observations, an appropriate **data filling** technique can be used to estimate an evenly spaced time series from the original observations which are available at irregular time intervals (see Section 19.3.2

for a discussion of data filling methods). For the situation where there is a large number of missing observations, a procedure based on **seasonal adjustment** can be used (see Section 22.2). As pointed out in Chapter 19, intervention analysis constitutes a very powerful technique for use in environmental impact studies.

An inherent advantage of using most nonparametric tests and regression analysis techniques given in Chapters 23 and 24, respectively, is that they can be used directly with evenly or unevenly spaced observations. As described in Chapter 23, since the early 1980's, a number of researchers have used **nonparametric tests** for detecting trends in water quality time series. A useful variety of nonparametric trend tests, including the Mann-Kendall and Spearman's partial rank correlation tests, are described in detail in Section 23.3. **Regression models** that can be employed as exploratory and confirmatory data analysis tools are discussed in Sections 24.2.2 and 24.2.3, respectively. The exploratory data analysis techniques described in Section 22.3 can, of course, be used in conjunction with the nonparametric tests and regression analysis, as well as the intervention models.

As summarized in Table 1.6.4, three **trend assessment methodologies** are presented in Part X for carrying out trend assessments of water quality and water quantity time series. For each of these studies, a methodological approach is developed within the overall framework of exploratory and confirmatory data analysis. The first study presented in Section 22.3 employs **intervention analysis** for modelling trends in water quality and water quantity time series measured in rivers. Within the second study discussed in Section 23.5, **nonparametric trend tests** and other appropriate statistical methods are utilized for discovering trends in water quality samples observed in a lake that may be affected by nearby industrial developments. Finally, in the third case study a methodology is designed in Section 24.3 for assessing trends in water quality time series measured in rivers. A particularly useful technique for tracing trends and accounting for the effects of flow upon a given water quality variable is the **robust locally weighted regression smooth** of Cleveland (1979) described in Section 24.2.2. Moreover, the **Spearman partial rank correlation test** is employed to detect trends in water quality time series when the impacts of seasonally are partialled out.

CHAPTER 22

EXPLORATORY DATA ANALYSIS AND INTERVENTION MODELLING IN CONFIRMATORY DATA ANALYSIS

22.1 INTRODUCTION

The main purpose of this chapter is to present a comprehensive methodology to identify and, if possible, stochastically model any trends which may be present in water quality as well as other kinds of environmental time series. These trends, if any, may be due to the presence of known or unknown interventions such as various types of land-use changes. In addition to possibly being affected by external interventions, usually a given water quality variable is measured at irregular time intervals and often there are large time gaps at which no data are collected. Consequently, water quality data are often very *messy* and systematic procedures are developed in this chapter, as well as Chapters 23 and 24, to optimize the amount of meaningful statistical information which can be gleaned from the currently available data.

As explained by Tukey (1977) and also briefly mentioned in Sections 1.2.4, 5.3.2, 19.2.3 and 24.2.2, there are usually two major steps in a statistical study. The first step is called *exploratory data analysis* and the objective of this phase of the work is to uncover important properties of the data by executing simple graphical and numerical studies. Some of the techniques available for this phase include a graph of the data against time, the 5-number summary graph which Tukey (1977, Ch. 2) calls the box-and-whisker plot, cross-correlation function, Tukey smoothing (Tukey, 1977, Ch. 7) and the autocorrelation function. The purpose of the next step which is referred to as *confirmatory data analysis* is to confirm statistically in a rigorous fashion the presence or absence of certain properties in the data. For example, when sufficient measurements have been taken for a water quality variable, exploratory data analysis may indicate that there is a possible trend in the data due to a known external intervention. Following this, the *intervention analysis* approach of Chapter 19 can be utilized as a confirmatory data analysis tool to determine if there has been a significant change in the mean level of the series.

The exploratory and confirmatory stages of data analysis can be compared to the process which takes place after a crime is committed (Tukey, 1977). At the exploratory stage of investigating a crime, a sleuth uses forensic tools and his common sense to discover evidence about the crime. If the detective does not understand how to execute an investigation, he may fail to look in the proper places for the criminal's fingerprints. On the other hand, if the investigator has no fingerprint powder he will not detect fingerprints on most objects. In an analogous fashion, the statistical analyst requires both the *tools of the trade* and *common sense*.

In the criminal justice system, the suspected criminal is taken to court after the collection of evidence by the investigative bodies. Following the evaluation of the available evidence, the jury and judge must ascertain if the criminal is guilty based upon the current information. Likewise, in a statistical study the purpose of the second main step, confirmatory data analysis, is to *verify quantitatively* if suspected statistical characteristics such as different kinds of trends are actually present in the data. When enough evidence is available, the results of a confirmatory data analysis can be quite useful to decision makers. For instance, when intervention analysis is employed in an environmental impact assessment study for properly confirming the presence of trends in water quality time series, the results can be used in court for forcing the polluters to adopt appropriate pollution abatement procedures.

Many exploratory and confirmatory methods require that equally spaced data be available, and as is pointed out earlier in this section, environmental series are often measured at uneven time intervals. Accordingly, in the next section a methodology based on *seasonal adjustment* is devised for estimating the entries of an average monthly time series when daily values are available at irregular time intervals and often there are time gaps spanning many months for which no measurements were taken. In addition to estimating values for a monthly sequence, this procedure can of course be used for estimating averages at other equal time intervals such as weekly or quarterly intervals by having fifty-two and four seasons per year, respectively.

Following the section on data filling, specific *exploratory data analysis techniques* are described in Section 22.3. In order to demonstrate clearly the efficacy of using exploratory data analysis and, when appropriate, the confirmatory data analysis tool of intervention analysis, *practical applications* are presented throughout the chapter. Possible *trends* in water quality and riverflow series are examined for two locations in Canada. In the province of Alberta, Canada, both exploratory and confirmatory data analysis techniques are employed to ascertain the effects of cutting down a forest upon total organic carbon and turbidity in the Cabin Creek near Seebe. On the Mill River near St. Anthony in Prince Edward Island, exploratory data analysis results suggest that perhaps due to acid rain, alkalinity levels may be increasing over time. These illustrative applications were originally presented in the paper by McLeod et al. (1983) and are in fact part of an extensive environmental study executed by the authors in which fifty environmental time series were exhaustively analyzed.

Besides the *intervention analysis* approach of Chapter 19 and Section 22.4 in this chapter, other confirmatory data analysis techniques include the *nonparametric tests* and *regression analysis* methods of Chapters 23 and 24, respectively. An advantage of most nonparametric tests and regression analyses is that they can be used with observations measured at either unequally or equally spaced time intervals. In fact, as pointed out in Chapter 23, nonparametric tests have been used extensively for checking for the presence of trends in water quality time series. In Section 23.3 and Appendices A23.1 to A23.3, many of these nonparametric tests are described in detail. As an alternative procedure, the regression models of Chapter 24, offer a promising parametric approach for modelling trends in unevenly spaced measurements.

It is important that the scientist always keep in mind the fact that *sufficient data* or information must be available if he or she wants to carry out confirmatory data analyses. Until further measurements are available, inadequate series must for the present time be "thrown out of court" due to lack of sufficient evidence. Often a detective has "a feeling" about whether or not a person is guilty of a crime. If he thinks that the suspect is actually guilty, he will continue to follow his prey until he collects sufficient information so the courts can eventually convict him. The same situation holds for statistical studies. Even though it may not be presently feasible to fit an intervention model or another type of confirmatory model to a specific water quality time series, if this series is deemed important, the further collection of data will eventually permit a full confirmatory data analysis study.

22.2 DATA FILLING USING SEASONAL ADJUSTMENT

Many exploratory data analysis methods are valid for use with either unequally or evenly spaced data. However, Tukey smoothing, which is explained in Section 22.3.5, is an example of an exploratory tool where the measurements, or estimates thereof, must be available at equal time intervals before the method should be used. Except for many of the nonparametric tests and also the regression analysis methods presented in Chapters 23 and 24, respectively, all of the stochastic models described in this book, including the intervention models of Chapter 19 and Section 22.4, can only be used with evenly spaced data at the confirmatory data analysis stage. Therefore, when data are unevenly spaced, procedures are required for creating an evenly spaced sequence which stochastically represents what could have occurred historically. As explained in Section 19.3, intervention analysis can be employed for estimating missing values from an evenly spaced data set when the number of unknown observations is not too large (usually not more than 5% of the data set). However, for evenly spaced daily observations with a large number of missing values, a different procedure must be adopted for estimating a sequence of evenly spaced average monthly values. The particular technique presented in this section is related to methods developed for seasonal adjustment models.

In *seasonal adjustment* modelling, a time series is decomposed into various components, one of which is the seasonal term. Various seasonal adjustment procedures are available and the reader may wish to refer to the statistical literature for a description of these techniques (see, for example, Kendall (1973), Shiskin et al. (1976), Granger (1980), Hillmer and Tiao (1985), and Cleveland et al. (1990)). Suppose that x_t represents an observation at time t either for the original time series or for some Box-Cox transformation of the given data. One reason for invoking the Box-Cox transformation in [3.4.30] is to cause data that are not normally distributed to approximately follow a normal distribution. For instance, a logarithmic transformation may reduce the skewness and improve the symmetry of the distribution if there are quite a few large values in the series. When the variance of a series depends on the level of the series, this transformation may rectify the problem. Furthermore, as explained in Section 3.4.5 and elsewhere, a Box-Cox transformation can often alleviate problems with the properties of the residuals of the stochastic model fitted to the series of equally spaced data.

An additive seasonal adjustment model can be written as:

$$x_t = C_t + S_t + I_t = C_r + S_m + I_t$$

where t is the Julian day number (i.e., the number of days since January 1, 4713 B.C.), r is the year, m is the month for monthly data, C_t or C_r is the trend factor for modelling relatively long term causes, S_t or S_m is a stable seasonal factor which is assumed not to evolve with time, I_t is the nonseasonal irregular component made up of short-run effects and is not necessarily white noise. The original seasonal adjustment algorithm presented by McLeod et al. (1983) consists of the following steps:

1. Obtain preliminary estimates of C_t, S_t and I_t. $\bar{C}_t = \bar{C}$ is taken to be a constant which is equal to the median of x_t. To get $\bar{S}_t$, first calculate $\bar{S}'_m$ as the median of $x_t - \bar{C}$ for the data in the mth month. Then use $\bar{S}_m = \bar{S}'_m - \frac{1}{12}\sum_{m=1}^{12}\bar{S}'_m$. Estimate the irregular component utilizing

$$\bar{I}_t = x_t - \bar{C} - \bar{S}_m$$

2. Replace far-out values in the $\bar{I}_t$ series by the nearest outer fence (see Section 22.3.3 on box-and-whisker graphs for definitions of far-out values and outer fences) to form the irregular series $\bar{I}'_t$. The process of replacing far-out values by outer fences is called *Winsorizing* (Tukey, 1977).

3. Estimate the deseasonalized series given by

$$D_t = \bar{C} + \bar{I}'_t$$

4. Determine the revised trend estimate, $\tilde{C}_t$, where each year in $\tilde{C}_t$ is the mean of D_t for that year. If no data are available for the rth year, the mean of D_t for surrounding years is used.

5. Calculate the revised seasonal component

$$\tilde{S}_m = \tilde{S}'_m - \frac{1}{12}\sum_{m=1}^{12}\tilde{S}'_m$$

where $\tilde{S}'_m$ is the median of $x_t - \bar{C}_t$.

6. The revised irregular series is estimated using

$$\tilde{I}_t = x_t - \tilde{S}_m - \tilde{C}_r$$

7. Winsorize the revised irregular series, $\tilde{I}_t$, to obtain the Winsorized series, $\tilde{I}'_t$. This is accomplished by replacing the far-out values of $\tilde{I}_t$ by the appropriate outer fences.

8. Obtain an adjusted version (i.e., Winsorized) of the x_t series using

$$x'_t = \tilde{C}_r + \tilde{S}_m + \tilde{I}'_t$$

For a given month for a specified year in which data were originally given, take the median of the x'_t values to get the estimated average monthly value.

9. Adjust the trend for each year by employing

$$\tilde{\tilde{C}}_r = \tilde{C}_r + \text{mean of } \tilde{I}'_t \text{ for the whole series.}$$

10. To obtain an estimated monthly average value for a given month in which no data were given use

$$\bar{x}_{r,m} = \tilde{\tilde{C}}_r + \tilde{S}_m$$

where $\bar{x}_{r,m}$ is the estimated monthly value for the rth year and mth month. The total estimated monthly series is formed by using Steps 8 and 10. Note that if a Box-Cox transformation is taken of the given data, then an inverse Box-Cox transformation must be invoked to obtain the estimated monthly averages for the original untransformed series.

In Section 24.2.2, the *robust locally weighted regression smooth (RLWRS)* of Cleveland (1979) is explained as a flexible procedure for smoothing a time series. The above seasonal adjustment algorithm can be improved by employing the RLWRS in the algorithm. In particular, the fourth step in the algorithm becomes:

4. Determine the revised trend estimate, $\tilde{C}_t$, as the RLWRS fitted to the deseasonalized series, D_t.

In order to demonstrate how well the seasonal adjustment algorithm works, consider the flows in m^3/s of the Cabin Creek near Seebe in Alberta, Canada, from January, 1964, till December, 1979. A daily flow value has been measured for each day during this time period and for each month in a given year an average monthly value can be readily calculated. Because riverflow measurements are often highly skewed, it is advantageous to take natural logarithms of the data. In Figure 22.2.1, the natural logarithms of the actual average monthly values are marked with black circles for one particular four year interval. For exactly the same days on which observations are missing for the turbidity data in the Cabin Creek, the corresponding daily observations are removed from the flow data. Following this, the seasonal adjustment algorithm is employed to estimate the average monthly flows of the logarithmic daily data for the period from 1964 to 1979 and these estimated flows are marked by circles in Figure 22.2.1. It should be pointed out that for the turbidity series and hence the estimated flows, only about 8% of the data are used in the seasonal adjustment algorithm. In addition, there are many months during which no observations are available. However, as can be seen in Figure 22.2.1, the estimated values from the seasonal adjustment algorithm are reasonably close to the actual entries during this four year period and also the other years not shown in Figure 22.2.1.

As noted in Section 22.1, the seasonal adjustment algorithm can also be used for estimating averages at equal time intervals other than monthly spacings. For instance, it can be employed for determining average bimonthly and quarterly time series. Moreover, the RLWRS discussed in Section 24.2.2 can be employed for improving step 4 of algorithm. The reader may wish to refer to Section 24.3.2 for a description of how the RLWRS can assist in analyzing trends of messy water quality series measured in rivers.

22.3 EXPLORATORY DATA ANALYSIS

22.3.1 Introduction

A wide range of exploratory data analysis tools are available for detecting important statistical characteristics contained in a data set (see, for example, Tukey (1977), Velleman and Hoaglin (1981), Berthouex et al. (1981), Chambers et al. (1983), Cluis (1983), McLeod et al. (1983), Hoaglin et al. (1983), du Toit et al. (1986) and Ramsey (1988)). In Section 19.2.3, a number of exploratory procedures are suggested for detecting known and unknown interventions in a time series and some of these techniques are discussed in detail in this section. Indeed, all of the identification tools which are recommended for designing the different kinds of time series models discussed throughout the book can be considered as exploratory data analysis techniques for specifically deciding upon the parameters required in these models. For example, in Figure 19.2.4, the graphical methods used to detect known and unknown interventions and the identification techniques needed to design the intervention models, could both be considered as exploratory data analysis tools. Nevertheless, in this section exploratory tools are presented which do not assume that a specific type of stochastic model will be used at the confirmatory data analysis stage. Because, in some situations, confirmatory data analysis may not be warranted, due, for example, to a lack of sufficient data, confirmatory data analysis may not be executed subsequent to exploratory data analysis. However, when a confirmatory data analysis is carried out, many of the results from the exploratory data analysis stage may be used along with specially designed

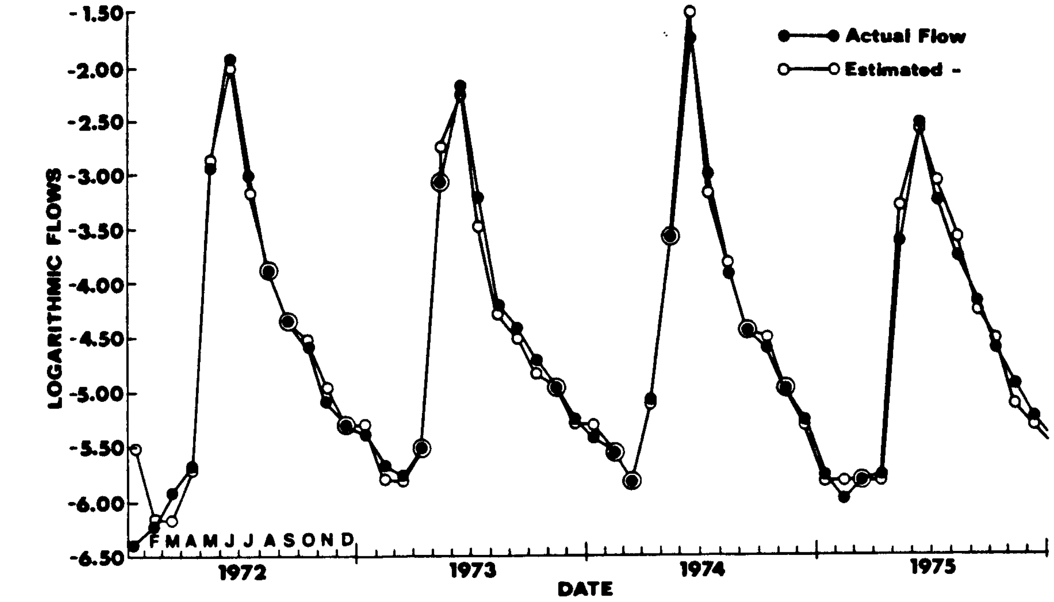

Figure 22.2.1. Monthly logarithmic flows of the Cabin Creek.

identification methods for constructing time series models at the confirmatory data analysis stage.

Some useful exploratory data analysis tools are presented in this section and the efficacy of using the techniques is demonstrated by applications to water quality and water quantity time series. Although many of the exploratory data analysis tools do not require the observations in a time series to be available at equally spaced time intervals, some of the techniques are designed for use with equally spaced measurements. The approaches discussed in this section which do not require equally spaced observations are:

1. graph of the data against time;
2. the 5-number summary graph which Tukey (1977, Ch. 2) calls a box-and-whisker plot;
3. cross-correlation function.

The following two techniques assume that the data points are separated by equal time intervals:

4. Tukey smoothing (Tukey, 1977, Ch. 7; Velleman and Hoaglin, 1981, Ch. 6);
5. autocorrelation function (ACF).

When data are unevenly spaced, an appropriate technique such as the seasonal adjustment algorithm from Section 22.2 or one of the other data filling methods of Section 19.2.3, can be employed for estimating the entries of an evenly spaced time series. Additionally, except for the third technique, all of the exploratory data analysis tools constitute valuable methods for detecting possible interventions.

The exploratory data analysis methods presented in this section and elsewhere can be employed for revealing interesting properties of the data under consideration. Each exploratory technique possesses its own inherent attributes that are useful for uncovering certain data characteristics. Because no single method can clearly portray everything there is to learn about the data, it is advantageous to examine the time series by employing a number of useful investigative graphical and numerical tools.

22.3.2 Time Series Plots

One of the simplest and more useful exploratory graphical tools is to plot the data against time. Characteristics of the data which may be easily discovered from a perusal of a graph include the detection of extreme values, trends, known and unknown interventions, dependencies among observations, seasonality, need for a data transformation, nonstationarity, and long term cycles.

When considering unequally spaced daily data, the actual time intervals between adjacent observations must be calculated before plotting the observations against time. A convenient technique to employ is to determine the *Julian day number* for each observation using the formula given by Hewlett-Packard (1977). With this information, the gap between adjacent observations can be determined as the difference of the Julian day numbers of the observations. This procedure is employed to obtain the graph in Figure 22.3.1 of the natural logarithms of the turbidity in the Cabin Creek, where each data point is marked by a small circle and is joined to its two neighbours by straight lines. As shown by the time gaps between observations, there are many days and even months during which no measurements were taken. For instance, from August 2 to November 22, 1975, inclusive, no observations were recorded.

Other examples of time series plots are presented throughout the textbook. In Chapter 2, Figure 2.3.1 displays the average annual flows of the St. Lawrence River at Ogdensburg, New York, for which there are no missing observations and no known interventions. An illustration of an annual series for which there is a known intervention is the average annual flows of the Nile River at Aswan in Figure 19.2.1. As seen in Figure 19.2.1 and discussed in Section 19.2.4, the completion of the Aswan dam in 1902 caused the average annual flows of the Nile River to decrease significantly from 1902 onwards. In Chapter 4, Figures 4.3.8, 4.3.10 and 4.3.15, show trends present in annual water use, electricity consumption, and Beveridge wheat price index time series, respectively. As noted in Section 4.3.3, each of these three nonstationary series can be adequately modelled using an ARIMA model.

Some interesting seasonal time series are also presented in the book. Consider, for example, Figure 1.1.1 or Figure 19.1.1 which displays the 72 average monthly phosphorous data (mg/l) from January 1972, until December, 1977, for measurements taken downstream from the Guelph sewage treatment plant located on the Speed River in the Grand River Basin, Ontario, Canada. As can be clearly seen, the commencement of phosphorous removal at upstream sewage treatment plants dramatically decreased the mean level of the series after the intervention date. Additionally, as indicated by the blackened circles, there are missing observations before and after the intervention. In Section 19.4.5, an intervention model is fitted to the water quality series in Figure 1.1.1 or 19.1.1 in order to statistically ascertain the effects of the intervention upon the level of the series and to estimate the missing observations.

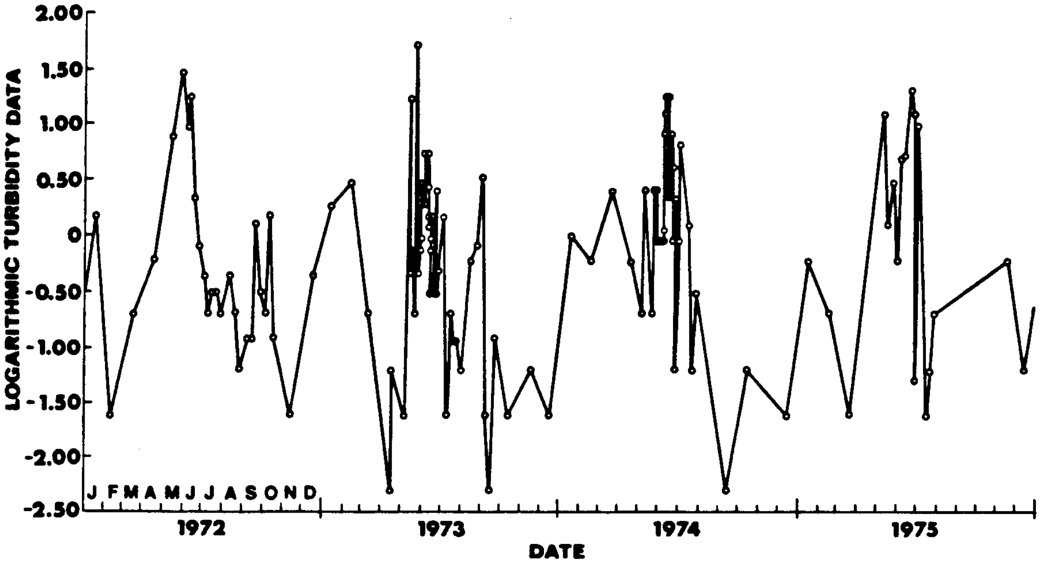

Figure 22.3.1. Natural logarithms of the turbidity (mg/l) data for the Cabin Creek.

22.3.3 Box-and-Whisker Graphs

A box-and-whisker graph is based upon what is called the *5-number summary* (Tukey, 1977, Ch. 2). For a given data set, the 5-number summary consists of the smallest and largest values, the median, and the 0.25 and 0.75 quantiles which are called *hinges*. When the data are ranked from the smallest to largest value, the first data point is the smallest value while the last entry is the largest value.

In order to calculate the values of *quantiles*, it is convenient to employ the operational definition of quantiles given by Chambers et al. (1983). Suppose that the given data represented by x_i for $i = 1,2, \ldots, n$, are ordered from smallest to largest such that the sorted data are denoted by $x_{(i)}$, $i = 1,2, \ldots, n$. If p represents any fraction between 0 and 1, the corresponding quantile is given by $Q(p)$. Whenever p is one of the fractions

$$p_i = (i - 0.5)/n \quad \text{for } i = 1,2, \ldots, n \qquad [22.3.1]$$

$Q(p)$ is assigned the value $x_{(i)}$, which is one of the given data points. For instance, if there were 10 observations, $x_{(2)}$ would have a p_i value of

$$p_2 = (2 - 0.5)/10 = 0.15$$

Hence, the 0.15 quantile, Q(0.15), would be exactly equal to $x_{(2)}$. When p is a fraction f of the way from p_i to p_{i+1}, one must use linear interpolation to estimate $Q(p)$. In particular, for this situation the interpolated quantile is calculated as

$$Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1}) \qquad [22.3.2]$$

Returning to the example for which there are 10 data points, the p_i for $x_{(3)}$ and $x_{(4)}$ are determined from [22.3.1] to be, respectively,

$$p_3 = (3 - 0.5)/10 = 0.25$$

$$p_4 = (4 - 0.5)/10 = 0.35$$

By utilizing [22.3.2], the quantile for $p = 0.31$ is determined to be

$$Q(0.31) = (1 - 0.6)Q(0.25) + 0.6Q(0.35)$$
$$= 0.4x_{(3)} + 0.6x_{(4)}$$

A plotting position is a value at which an ordered observation in a sample should be plotted for use on probability paper. In [22.3.1], p_i stands for the plotting position of the ith ordered value denoted by $x_{(i)}$. The plotting position given by [22.3.1] is actually a special case of the general plotting position formula written as

$$p_i = (i - \alpha)/(n + 1 - 2\alpha) \qquad [22.3.3]$$

where α is usually assigned values between 0 and 0.5. Detailed discussions regarding probability plotting positions are given by Barnett (1975) and Cunnane (1978) within the statistical and hydrological literature, respectively. As explained by these authors, when $\alpha = 0$ in [22.3.3] one obtains Weibull's formula which is recommended for use with uniformly distributed data. For normal observations, Blom's formula using $\alpha = 3/8$ in [22.3.3] should be employed. Finally, [22.3.1] is referred to as Hazen's formula and to obtain this, one substitutes $\alpha = 0.5$ into [22.3.3].

As noted by Chambers et al. (1983), there are many reasons for choosing p_i to be $(i - 0.5)/n$ in [22.3.1] or, equivalently, $\alpha = 0.5$ in [22.3.3], rather than some other value such as i/n. One consideration is that when the ordered observations are split into two groups exactly on an observation, the use of $(i - 0.5)/n$ means that the observation is counted as being half in the lower group and half in the upper group.

Because extrapolation must be done only when necessary and with great care, the formula in [22.3.2] for calculating $Q(p)$ should not be used outside the range of the data for which p is smaller than $0.5/n$ or larger than $1 - 0.5/n$. The safest rule for extrapolation is to define $Q(p) = x_{(1)}$ for $p < p_1$ and $Q(p) = x_{(n)}$ for $p > p_n$. According to this rule, $Q(0)$ and $Q(1)$ are assigned values of $x_{(1)}$ and $x_{(n)}$, respectively, which are the smallest and largest observations, respectively, in the given data set.

Equations [22.3.1] and [22.3.2] can be used with a data set of length $n \geq 2$. Keeping in mind that the only difference between a *percentile* and a quantile is that a percentile refers to a percentage of a data set and a quantile refers to a fraction of the data, these equations can also be used to calculate a percentile.

The *median*, given by $Q(0.5)$, divides the data into two groups of equal size. If n is odd, the median is $x_{((n+1)/2)}$. When n is even, $Q(0.5)$ is calculated using [22.3.2] as the average of $x_{(n/2)}$ and $x_{(n/2+1)}$, which are the two ordered values closest to the middle.

The lower and upper quartiles which are defined as $Q(0.25)$ and $Q(0.75)$, respectively, are called *hinges* by Tukey (1977). The distance between the first and third quantile, given by $Q(0.75) - Q(0.25)$ is called the *interquartile range*. This distance, which can be used to judge the spread of the data, is referred to by Tukey (1977) as the *H spread*.

To assist in characterizing extreme values, Tukey (1977) has suggested the following definitions. A *step* is 1.5 times the H-spread or interquartile range. *Inner fences* are one step outside hinges and *outer fences* are two steps outside hinges. Values between an inner fence and its neighbouring outer fence are called *outside*. Values beyond outer fences are *far-out*. Assuming that the data follow a given distribution, such as a normal distribution, one can calculate the expected numbers of outside and also far-out values, and compare these to the observed numbers.

When entertaining *seasonal data* such as monthly or quarterly data, it is instructive to calculate a 5-number summary plus outside and far-out values for each season. A convenient manner in which to display this information is to plot a box-and-whisker diagram for each season. Figure 22.3.2 depicts the monthly box-and-whisker plots for turbidity in the Cabin Creek before July 1, 1974, when part of the forest was cut down. In this figure, the data have not been transformed using a Box-Cox transformation. The upper and lower ends of a rectangle for a given month represent the two hinges and the thick line drawn horizontally within the rectangle is the value of the median. The minimum and maximum values for a particular month are the end points of the lines or *whiskers* attached to the rectangle or *box*. The far-out values are indicated by a circle in Figure 22.3.2, where far-out values are not marked if there are four or less data points for a given month. Below each month is a number which gives the number of data points used to calculate the box-and-whisker graph above the month. When there are not many data points used to determine a box-and-whisker plot for a given month, any peculiarities in the plot should be cautiously considered. The total number of observations across all the months is listed below November and December.

Another way to investigate *extreme values* is to calculate far-out values when all of the data across all of the seasons are used. Certainly, if a data point is far-out overall, the scientist should determine whether the measurement is accurate and represents what actually occurred or the observation is really due to measurement error or some other type of mistake. If the validity of a far-out overall or perhaps a far-out seasonal value is in doubt, in certain situations it may be advantageous not to include this data point in subsequent analyses. In the data filling procedure described in Section 22.2, far-out values are adjusted using a technique called *Winsorizing* (Tukey, 1977).

In addition to detecting far-out values, box-and-whisker diagrams have other uses. If the data are approximately symmetrical with respect to the median, they may follow a *symmetric distribution* such as the normal distribution. If there is an obvious lack of symmetry in a box-and-whisker graph, this may indicate the need for a transformation, such as the Box-Cox transformation in [3.4.30] to cause the data to be approximately normally distributed. Since, by definition, 25% of the data is contained between the median and a hinge, for normally distributed data the hinge is located 0.68 times the standard deviation from the median or mean. It can also be shown for normal data that inner fences are located a distance of 2.70 standard deviations on either side of the mean and the outer fences are a distance of 4.72 standard deviations from the mean. Consequently, the probability of having a far-out value with normally distributed data, is extremely small. Therefore, using a transformation to normalize a given data set will tend to

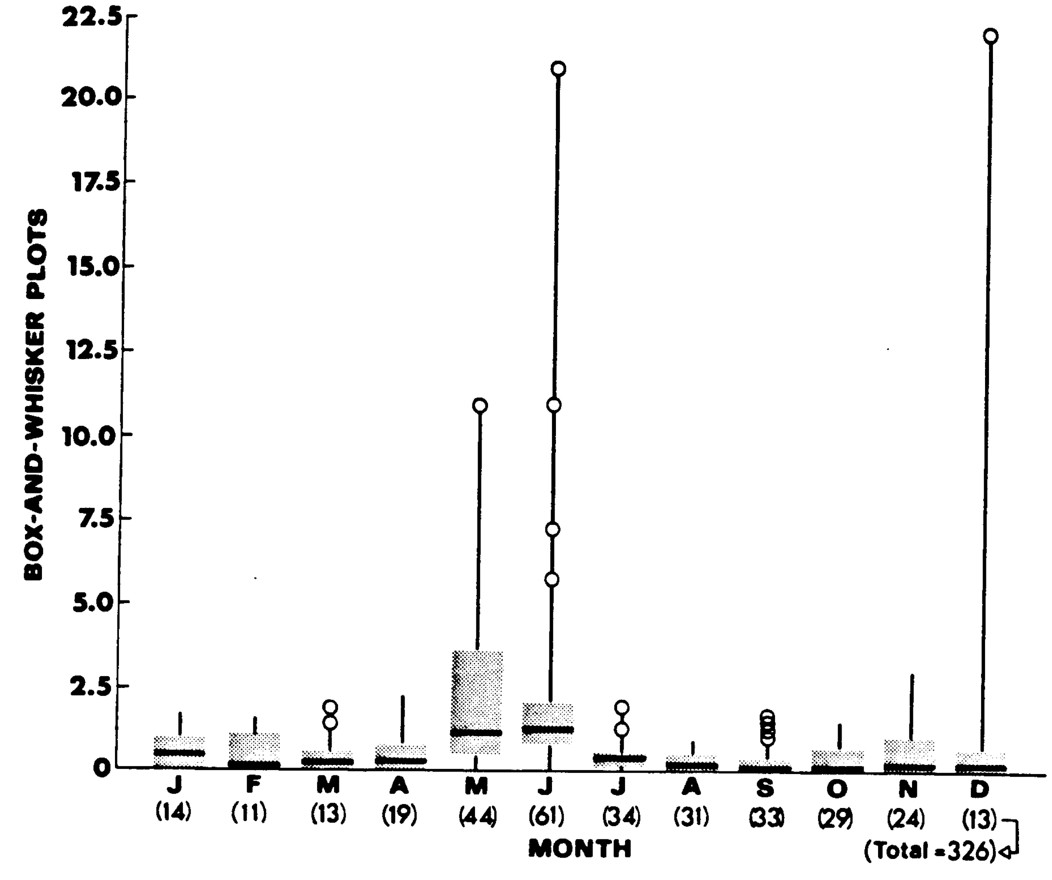

Figure 22.3.2. Box-and-whisker plots for turbidity (mg/l) in the Cabin Creek before July 1, 1974, where there is no data transformation.

reduce the number of far-out values.

For a given season in a box-and-whisker diagram, *symmetric data* would cause the median to lie in the middle of the rectangle and the lengths of the upper and lower whiskers would be about the same. Notice in Figure 22.3.2 for the turbidity data that the whiskers are almost entirely above the rectangle for all of the months and for six of the months there are a total of 14 far-out values. This lack of symmetry can at least be partially rectified by transforming the given data using the Box-Cox transformation in [3.4.30]. By comparing Figure 22.3.2 to Figure 22.3.3 where natural logarithms are taken of the turbidity data, the improvement in symmetry can be clearly seen. Furthermore, the Box-Cox transformation has reduced the number of far-out entries from 14 in Figure 22.3.2 to three in Figure 22.3.3.

Box-and-whisker plots can be employed as an important exploratory data analysis tool in *intervention studies*. If the date of the intervention is known, box-and-whisker diagrams can be constructed for each season for the data before and after the time of intervention. These two

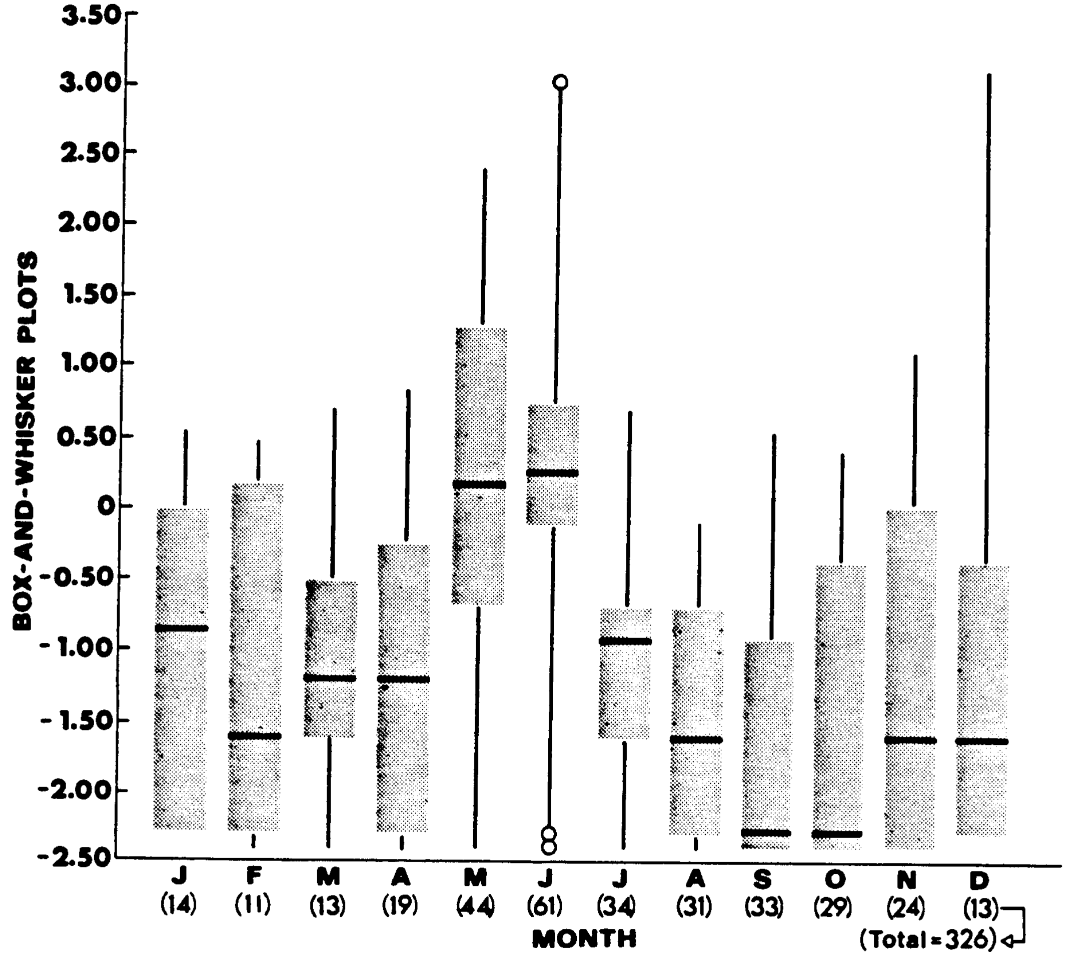

Figure 22.3.3. Box-and-whisker plots for turbidity (mg/l) in the Cabin Creek before July 1, 1974, where there is a logarithmic data transformation.

graphs can be compared to ascertain for which seasons the intervention has caused noticeable changes. When there are sufficient data, this type of information is crucial for designing a proper intervention model to fit the data at the confirmatory data analysis stage.

The Cabin Creek basin, which has an area of 2.12 km^2, was originally forested but from July to October, 1974, 40% of the forested area was clear-cut. Total organic carbon readings are available from March 17, 1971, to January 10, 1979. Figures 22.3.4 and 22.3.5 display the box-and-whisker plots of the natural logarithms of the total organic carbon in mg/l for the Cabin Creek before and after the intervention, respectively, caused by the removal of the trees. As can be observed, there are obvious drops in the medians for almost all the months after the intervention. These and other changes cannot be as easily detected in a plot of the entire series against time.

When examining seasonal box-and-whisker diagrams, such as those given in Figures 22.3.2 to 22.3.5, one may wish to compare the statistical characteristics of data among seasons in order to ascertain if there are any significant differences. One may be tempted, for instance, to check if two boxes do not overlap with one another which is the case for the June and July box-and-whisker plots in Figure 22.3.3. Unfortunately, the hinges which delineate the top and bottom of each box are not the appropriate guides to employ when checking for significant differences in a statistic such as the median between two seasons. A way for comparing medians between two box-and-whisker diagrams to check if they are significantly different, is to employ the *notched box-and-whisker plot* concept proposed by McGill et al. (1978). More particularly, for each box-and-whisker plot, a notch of specified size given below is drawn on the left and right side of the box with its centre at the median. When two box-and-whisker plots are compared to one another, the two medians are significantly different or not different according to whether the notches do not overlap or overlap, respectively. A suggested size for the notch is the

$$\text{median} \pm 1.58 \times \text{H–spread}/\sqrt{n}$$

where n is the number of data points used to construct the box-and-whisker plot for a given season. Assuming normality and independence of the data for each season, the significance level for which the median test is designed is approximately the 5% level. Examples of notched seasonal box-and-whisker plots are given in Figure 24.3.4 in Section 24.3.2.

Notice that at the bottom of each month for the seasonal box-and-whisker plots drawn in Figures 22.3.2 to 22.3.5, the number of data points is given. An approach for visually portraying the number of observations is to make the width of a box to be proportioned to the number of measurements. McGill et al. (1978) suggests drawing *variable-width box-and-whisker plot* for which the width is proportional to $\sqrt{n}$ for each season or group of data.

22.3.4 Cross-Correlation Function

In Section 16.2, it is explained how meaningful causality can be detected between two series labelled x_t and y_t, when the observations in each time series are equally spaced and sufficient observations are available. Subsequent to fitting an appropriate ARMA model to each of the series, the sample cross-correlation function (CCF) between the estimated residuals or prewhitened series of the two data sets can be calculated using [16.2.6]. By examining the properties of the residual CCF at negative, zero and positive lags, the type of causality between x_t and y_t can be ascertained. Illustrative applications for using this procedure are given in Section 16.3 and, in Sections 17.3.1, it is explained how the information from the residual CCF analysis can be used for identifying a TFN model to link the x_t and y_t series, when this type of model is warranted. In Sections 20.3.2 and 21.3.2, it is described how the residual CCF can be used to identify when a general multivariate ARMA model and a CARMA (contemporaneous ARMA) model are needed to model formally the mathematical relationship between x_t and y_t.

In exploratory data analysis, often there may be many missing data points and before two series can be prewhitened by fitting an ARMA model to each series, evenly spaced time series must be estimated. Further, at the exploratory data analysis stage one may only wish to have a general idea about the relationship between two series and to examine the CCF of the two given series at lag zero. If necessary, at a later stage in the data analysis study a proper residual CCF analysis can be executed. Consequently, even before an evenly spaced series is estimated for the

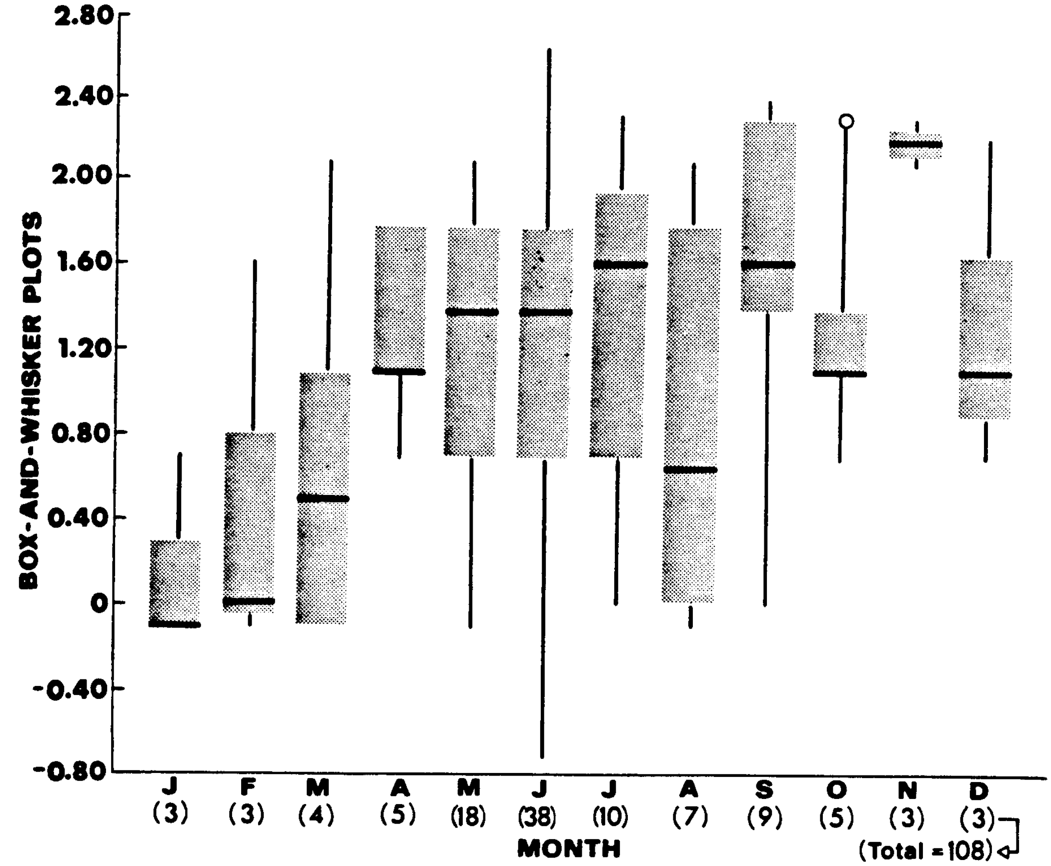

Figure 22.3.4. Box-and-whisker plots of the logarithmic total organic carbon (mg/l) in the Cabin Creek before July 1, 1974.

situation where data are missing, the CCF for x_t and y_t can be calculated for the values of the two series which are measured at the same time.

The CCF between two time series can be calculated to determine the amount of linear dependence between the two series. When x_t represents the observation recorded at time t for one series and y_t is the observed value at the same time for a second series, the *sample CCF* at lag zero can be calculated using

$$r_{xy}(0) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\left[\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2\right]^{1/2}} \qquad [22.3.4]$$

where n is the number of times observations occur at the same time in the two series, $\bar{x}$ is the mean of the x_t series, and $\bar{y}$ is the mean of the y_t series. The value of $r_{xy}(0)$ can range from -1 to

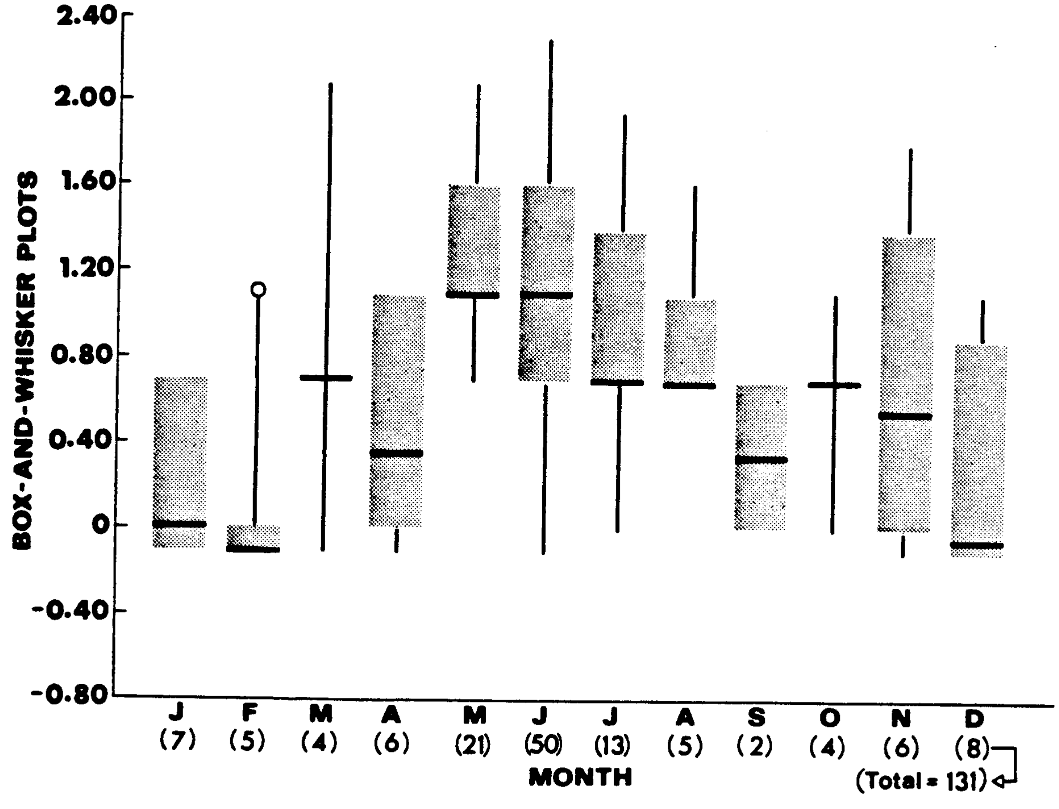

Figure 22.3.5. Box-and-whisker plots of the logarithmic total organic carbon (mg/l) in the Cabin Creek after October 31, 1974.

+1. If the x_t and y_t series are both white noise and also independent of one another, for large samples $r_{xy}(0)$ is normally independently distributed with a mean of zero and variance of $1/n$ (Haugh, 1976). Consequently, the 95% confidence limits are approximately $\pm 1.96n^{-1/2}$.

The sample CCF at lag zero can be calculated for either the original series or else the series transformed using the Box-Cox transformation in [3.4.30]. Recall that when the parameter $\lambda = 1$ in [3.4.30], this indicates that there is no transformation while $\lambda = 0$ means that each data point is transformed using natural logarithms. Suppose that a number of water quality variables plus riverflows have been measured at one location in a river. For a given site, $r_{xy}(0)$ in [22.3.4] can be calculated for all possible pairs of water quality and water quantity time series. Consider, for example, seven time series measured on the Mill River near St. Anthony, Prince Edward Island, Canada. Table 22.3.1 lists according to a number each of the seven series where the Box-Cox transformation in [3.4.30] used for each series is given. Below the list of series is the correlation matrix which is calculated using [22.3.4]. An (i,j) entry in the correlation matrix gives the value of the CCF at lag zero between series i and j which are defined above the correlation matrix in the table. For example, in Table 22.3.1 the CCF at lag zero between the 6th and 2nd series is

-0.817. Because this is the same as $r_{xy}(0)$ between series 2 and 6, the correlation matrix is symmetric and only the lower part of the matrix is presented. Moreover, the negative value indicates that the observations in one series tends to be larger whenever the values in the other series are smaller, and vice versa. Notice that all the diagonal entries have a value of unity since a series is fully correlated with itself at lag zero. If an $r_{xy}(0)$ value is within the range of $\pm 1.96n^{-1/2}$ it is automatically assigned a value of zero to indicate that it is not significantly different from zero.

Table 22.3.1. Cross-correlations for the Mill River time series.

DATA SETS

1. pH (pH Units), $\lambda = 1$
2. Stability Index, $\lambda = 0$
3. Daily Mean Discharge (m^3/s), $\lambda = 0$
4. Dissolved Sulphate (mg/l), $\lambda = 0$
5. Total Alkalinity (mg/l), $\lambda = 0$
6. Dissolved Calcium (mg/l), $\lambda = 1$
7. Water Temp. (degrees Celcius), $\lambda = 1$

CROSS-CORRELATION MATRIX

1.000						
-0.908	1.000					
-0.764	0.916	1.000				
0.268	-0.397	-0.436	1.000			
0.827	-0.965	-0.929	0.321	1.000		
0.535	-0.817	-0.793	0.000	0.768	1.000	
0.326	-0.376	-0.402	0.000	0.366	0.316	1.000

Measuring a large number of phenomena at a given location is usually quite expensive. If it is required to record less variables in order, for example, to allow the remaining items to be measured more frequently, the cross-correlation matrix may be helpful for deciding upon which variables to continue measuring. When one variable is highly correlated with another, then measuring only one of the variables furnishes an indication of the possible magnitudes of the actual values for the other unobserved variable. Consequently, based upon a firm understanding of the actual physical process plus the statistical evidence in the cross correlation matrix, it may be feasible to only continue to measure one of the series. If enough equally spaced measurements are taken of the remaining variable to permit the resulting time series to be thoroughly studied using a technique such as intervention analysis (see Chapter 19) at the confirmatory data analysis stage, this could be of great benefit to the decision makers.

A perusal of Table 22.3.1 reveals that many variables are highly correlated with one another. For example, notice in Table 22.3.1 for the Mill River time series that the stability index is highly correlated with pH($r_{2,1}(0) = -0.908$), daily mean discharge ($r_{3,2}(0) = 0.916$), total alkalinity ($r_{5,2}(0) = -0.965$), and dissolved calcium ($r_{6,2}(0) = -0.817$). Of course, it is known from a definition of the stability index that it is a function of the other mentioned water quality variables and this is confirmed by the appropriate entries in the correlation matrix in Table 22.3.1.

The zero entries in Table 22.3.1 demonstrate that sometimes there is no significant linear dependence between many of the variables. For example, in Table 22.3.1 for the Mill River series, the value for $r_{7,4}(0)$ between water temperature and dissolved sulphate is not significantly different from zero. When no significant cross-correlation exists between two series, then the decision about possibly dropping one of the series must be based upon other factors.

22.3.5 Tukey Smoothing

Introduction

Sometimes a graph of a given time series *blurs* statistical information in the data which a smoothed plot of the series at equally spaced time intervals may reveal more clearly. Consider, for example, Figure 22.3.6, which is a plot of the average annual total organic carbon in mg/l, for the Cabin Creek where the average annual entries are calculated using the estimated monthly values obtained from the seasonal adjustment algorithm developed in Section 22.2. In this graph, there appears to be a drop in the mean level of the series in the later years compared with the values in the early 1970's. When the *blurred smooth* in Figure 22.3.7 is studied, the general characteristics of the data are more clearly portrayed. Figure 22.3.7 is a blurred smoothed plot of the average annual total organic carbon for the Cabin Creek where the vertical lines reflect the magnitude of the rough or blur of the series and a smoothed observation is located at the midpoint of the bar. Notice from Figure 22.3.7 that the smoothing characteristics for the data before 1974 are more or less the same but from 1974 onwards there is an obvious decrease in the mean of the series. This property was also suggested by the box-and-whisker plots of the series shown before and after the intervention in Figures 22.3.4 and 22.3.5, respectively.

Although a *smoothed graph* does not contain any more information than what is already present in the plot of the raw data, in many instances the smoothed graph portrays the essential features much more clearly. The purpose of a smoothed curve is to reveal the systematic structure and interesting statistical characteristics of the data. Consider, for example, the blurred smoothed graph in Figure 22.3.8 for the total alkalinity in mg/l for the Mill River at St. Anthony in Prince Edward Island, Canada. This graph is a blurred smoothed plot of the average annual values which were calculated from the estimated monthly entries obtained from the seasonal adjustment algorithm in Section 22.2. In Figure 22.3.8, there is an obvious shift downwards in alkalinity from 1973 to 1977 followed by abrupt decreases in 1978 and 1979. Because the soil in the Mill River basin is sandy, acid rain could quickly drain through the ground without undergoing substantial chemical changes and thereby adversely affect the water quality. Consequently, the decrease in alkalinity in Figure 22.3.8 could be mainly due to acid rain which could severely affect the biological life in the river. However, it is still necessary to collect more data and determine when the acid rain intervention came into effect before proper confirmatory data analyses can be executed.

To construct a smoothed curve, consider qualitatively subdividing a given time series as

$$Data = Smooth + Rough$$

By filtering out the *rough* or noise portion of the data, the smoothed curve or *smooth* can be examined for important statistical features. The filter which maps the given series into a

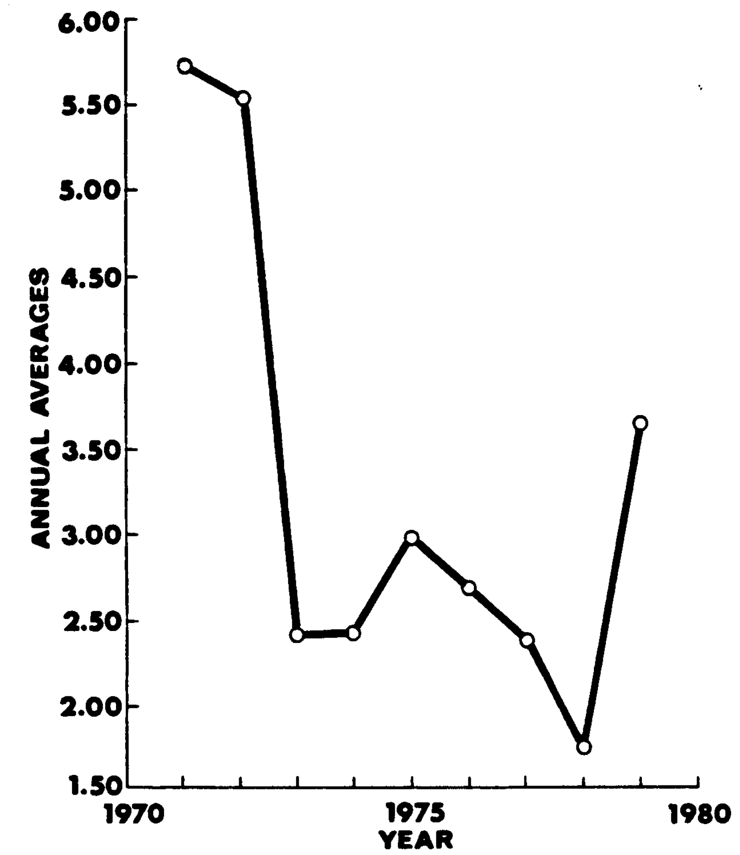

Figure 22.3.6. Estimated annual values of the total organic carbon (mg/l) in the Cabin Creek.

smoothed curve is referred to as a *smoother*. A trace plot of the smooth (against time) can display trends and changes in level of the series more clearly than a plot of the raw data. Of equal importance, a graph of the rough (over time) can reveal outliers, changes in variance and other unusual features.

Smoothed curves could be calculated for time series available at different time intervals between each pair of data points such as daily, monthly or yearly time separations. If one is attempting to detect *short term trends*, then it may be advantageous to examine smoothed curves and also time series plots for data points separated by short time intervals. However, at short time intervals long term trends may not be as easy to detect due to the large amount of rough in the data. Consequently, in order to discover *long term trends*, it may be advantageous to use annual data as is done in Figure 22.3.6 for the time series plot of the estimated annual values of total organic carbon (mg/l) in the Cabin Creek and Figure 22.3.7 of the blurred 3RSR smooth of the annual data plotted in Figure 22.3.6. As noted earlier, the long term trend in the total organic carbon time series can be easily visualized by examining these two figures. Within this section

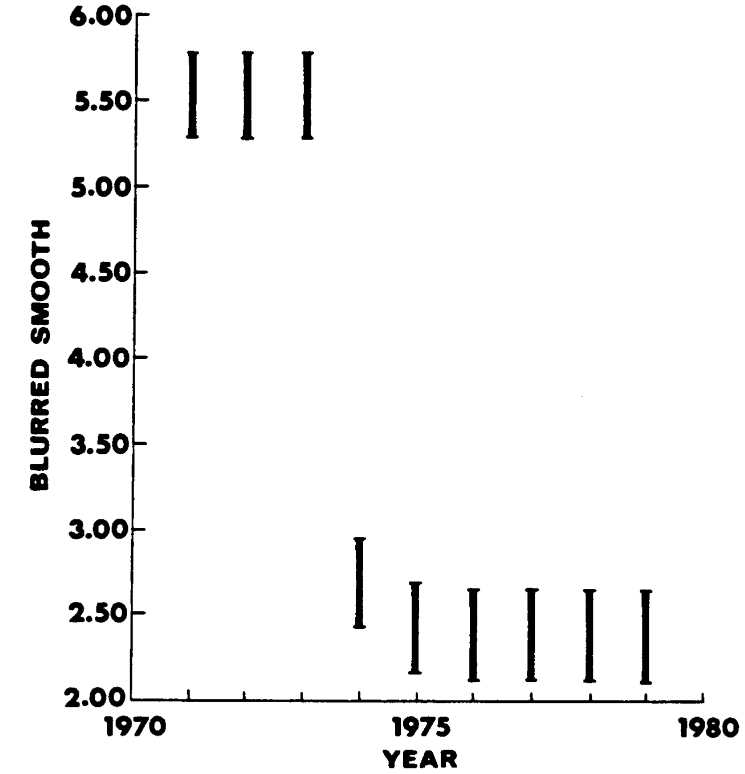

Figure 22.3.7. Blurred 3RSR smooth of the estimated average annual total organic carbon (mg/l) in the Cabin Creek.

and also Section 22.3.6, annual time series are considered so that long term trends can be conveniently discovered and their behaviour can be better understood.

The nonlinear smoothers developed by Tukey (1977, Ch. 7) and also discussed by McNeil (1977), are very flexible when used in practical applications and are capable of detecting all of the items discussed for a plot of the series except, possibly, for occasional outliers. Mallows (1980) explains the desirable properties that any smoother should possess and also presents some theoretical mathematical results for Tukey smoothers. Some of the more important attributes that a smoother should have include the ability to be responsive to abrupt changes in level, marginal distribution, and covariance structure.

Figures 22.3.7 and 22.3.8 are examples of what Tukey (1977, Ch. 7) calls a *blurred 3RSR smooth*. In fact, Tukey defines a variety of useful nonlinear smoothers. Within the next subsection the blurred 3RSR smooth is defined while the *4253H, twice, smooth* is described in the last part of Section 22.3.5. The reader may also wish to read about the flexible smooth of Cleveland (1979) which is based upon regression analysis and defined and applied in Sections 24.2.2 and 24.3.2, respectively. Finally, Velleman (1980) provides comparisons of robust nonlinear data smoothing algorithms.

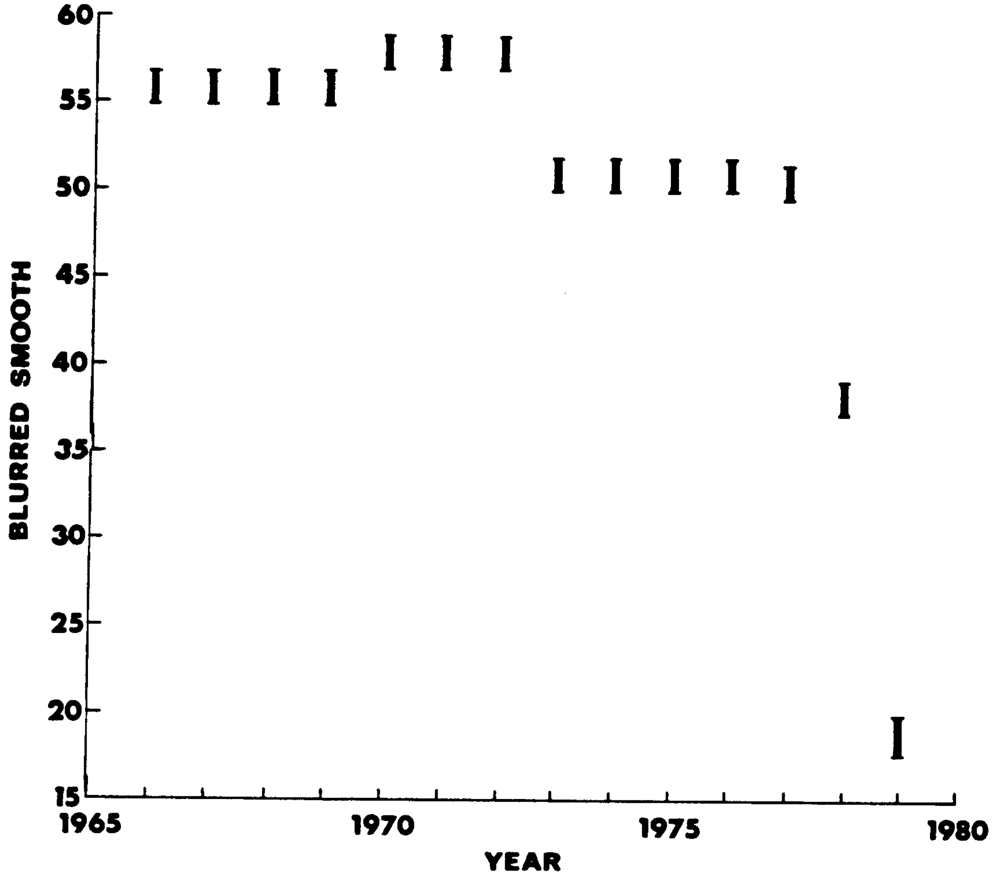

Figure 22.3.8. Blurred 3RSR smooth of the estimated average annual total alkalinity (mg/l) in the Mill River.

Blurred 3RSR Smooth

Blurred 3RSR smooths are displayed in Figures 22.3.7 and 22.3.8. When developing a *blurred 3RSR smooth* for a series $x_1, x_2, \ldots, x_n$, various calculations must be done (Tukey, 1977, Ch. 7; McNeil, 1977, Ch. 6) and these are outlined below.

1. *Smoothing by repeated medians of 3 (called 3R)* - To smooth the given time series using running medians of 3, replace the observation at time t by the median of x_{t-1}, x_t, and x_{t+1}. The smoothed values for the end points x_1 and x_n are calculated according to the rules in Step 2. Next, the smooth which was just determined is itself smoothed by using running medians of 3 and once again the smoothed end points are determined from Step 2. This procedure is repeated until the curve can be smoothed no further. The label 3R indicates that the smoothing is repeated using running medians of 3 until convergence is reached.

2. *Smoothing end points* - To obtain the smoothed values for the end points of the given series (or for the end points to any given sequence) replace x_1 by the median of x_1, x_2 and $3x_2 - 2x_3$ and substitute the median of x_n, x_{n-1}, and $3x_{n-1} - 2x_{n-2}$ for x_n.

3. *Repeated splitting (called SR)* - The use of medians instead of means tends to create mesas which are pairs of adjacent points with the same value that is below or above the points at each side of the mesa. Therefore, a *mesa* is simply a two-point local minimum or maximum. To smooth the mesas, split the series at the centre of each mesa and apply the end-point rule in Step 2 separately to the values on each side of each of the divisions. The resulting series is then smoothed using the 3R method in Step 1. The procedure of using the end-point rule for replacing the values at the mesas and then employing 3R is called splitting (i.e., S). If mesas still exist after splitting, the splitting is repeated until either all of the mesas disappear or convergence is reached. This is referred to as repeated splitting or simply SR. A computer program for calculating the 3RSR curve, which is created upon the completion of Step 3, is given by McNeil (1977, Ch. 6).

4. *Blurring* - When Steps 1 to 3 are used to obtain the smooth 3RSR a series of single points can be plotted. To reflect the variation, blur or rough in the series beyond the 3RSR curve, vertical bars can be plotted which are centered at each point on the 3RSR smooth. To calculate the length of the bars, first determine

   rough = data − smooth (i.e. 3RSR)

   for each point in the series. Wherever the rough has a value of exactly zero replace it by 0.5 as was suggested by Tukey (1977, Ch. 7). The bar length is then taken as the magnitude of the median of the absolute values of all the roughs. The 3RSR curve which is plotted using bars is called the blurred 3RSR smooth.

Summary - To obtain a blurred 3RSR smooth first determine the smooth 3R from Step 1 where the smoothed end points are calculated using Step 2. Next, determine the curve 3RSR by employing repeated splitting according to Step 3 and the relevant portions of Steps 1 and 2. Finally, Step 4 can be utilized to procure a blurred 3RSR smooth. Figures 22.3.7 and 22.3.8 are examples of blurred 3RSR smoothes which are calculated using the foregoing algorithm.

4253H, Twice Smooth

A particularly robust smooth is the *4253H, twice smooth* (Velleman and Hoaglin, 1981, Ch. 6). As indicated by the name, it involves taking medians of 4, then 2, then 5, then 3, then Hanning and then applying 4253H to the residuals of the first pass and adding this to the first pass smoother. To clarify how each step is calculated, an illustrative example is included in the explanation given below.

1. *Smooth using running medians of 4* - When determining the median of an even number of observations, the measurements being considered are ranked from smallest to largest and the median is taken as the average of the middle two values. Consequently, when calculating the median of four observations, the four values are listed in ascending order of magnitude, and the median is the average of the two middle numbers. For smoothing in using running medians of 4, which is simply called 4 smoothing, the endpoints themselves are just *copied on* in the 4 smooth series. The value located second from either end in the 4 smooth is simply the averages of the appropriate two end points in the given series. All

other entries in the 4 smooth are calculated as running median of 4.

As an illustrative example which is used to explain the entire 4253H, twice smooth, consider a hypothetical sequence of eight values given as:

5, 2, 4, 4, 0, 2, 3, 4.

The 4 smooth consisting of nine values is found to be:

5, 3.5, 4, 3, 3, 2.5, 2.5, 3.5, 4.

To explain how each value in the 4 smooth is determined, the calculations are given below for each of the numbers by starting on the left and working to the right.

5 = *copied on*

3.5 = *med*(5,2) = (5+2)/2 = 3.5

4 = *med*(5,2,4,4) = (4 + 4)/2 = 4

3 = *med*(2,4,4,0) = (2 + 4)/2 = 3

3 = *med*(4,4,0,2) = (2 + 4)/2 = 3

2.5 = *med*(4,0,2,3) = (2 + 3)/2 = 2.5

2.5 = *med*(0,2,3,4) = (2 + 3)/2 = 2.5

3.5 = *med*(3,4) = (3 + 4)/2 = 3.5

4 = *copied on*

2. *Smooth utilizing running medians of 2* - In step 1, except for the end points, each sequence of two numbers in the smooth can be thought of as lying on either side of the appropriate number in the original series in terms of the time axis. For example, the second and third entries from the left in a 4 smooth can be interpreted as residing on both sides of the second number in the original series. To line up in time the 4 smooth with the given series, a running median of two is applied to the 4 smooth after copying on the end points. In other words, the average is calculated for each sequential set of two values in the 4 smooth, excluding the two end points.

   In terms of the application, applying a 2 smooth to the previous 4 smooth creates the series of eight values given as:

   5, 3.75, 3.5, 3, 2.75, 2.5, 3, 4.

   Notice that the end points given by 5 and 4 are simply copied on. The second number from the left is calculated as:

   3.75 = *med*(3.5,4) = (3.5 + 4)/2.

3. *Smooth using running medians of 5* - When determining the median of an odd number sequence of observations, the measurements being entertained are ranked from smallest to largest and the median is selected as the value falling in the middle. For a 5 smooth, the ends are just copied on and lower order smoothing is employed near the ends while running medians of 5 are employed to calculate all other entries.

When the 5 smooth is applied to the smooth obtained at step 2, the result is

5, 3.75, 3.5, 3, 3, 3, 3, 4.

Once again the two end points are copied on as 5 and 4. The second entry from the left is determined as:

$$3.75 = med(5, 3.75, 3.5).$$

Likewise, the second value from the right is found to be:

$$3 = med(2.5, 3, 4).$$

All other values are determined using running medians of 5. For instance, the third entry from the left is found using:

$$3.5 = med(5, 3.75, 3.5, 3, 2.75).$$

4. *Smoothing employing running medians of 3* - Suppose that the series to be smoothed at this stage is represented as $x'_1,x'_2, \ldots, x'_n$, where the original series is represented as $x_1,x_2, \ldots, x_n$. The left and right end points of the 3 smooth are calculated as:

$$med(3x'_2 - 2x'_3, x'_1, x'_2) \text{ and}$$

$$med(x'_{n-1}, x'_n, 3x'_{n-1} - 2x'_{n-2}), \text{ respectively.}$$

Notice that the point $3x'_2 - 2x'_3$ corresponds to an estimate of x'_0 derived by extrapolating backwards the line joining $(2, x'_2)$ and $(3, x'_3)$. By working from left to right, all other entries, excluding the end points, are calculated as running medians of 3.

Applying the 3 smooth to the example as developed in the previous steps, produces the series

4.25, 3.75, 3.5, 3, 3, 3, 3, 3.

The left end point is determined using:

$$4.25 = med(3(3.75) - 2(3.5),5,4)$$
$$= med(4.25,5,4).$$

The second entry from the left is found as:

$$3.75 = med(5,3.75,3.5)$$

5. *Hanning* - Hanning (H) refers to a rather gentle smoothing operation for which a given value is replaced by a ***running weighted average***. Let the current series to which hanning is to be applied be given as $(x_1'', x_2'', \ldots, x_n'')$. A practical running weighted average to employ for calculating the hanned value at time t is

$$\frac{1}{4}x_{t-1}'' + \frac{1}{2}x_t'' + \frac{1}{4}x_{t+1}''.$$

As can be seen, the weights given by $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ sum to unity. Although an unlimited number of running weight averages are available for use, the hanning smoother utilized here employs the weights given above. When employing hanning, the two end points are copied on and all other entries are calculated using the running weighted average just presented.

For the case study as developed in the previous step, applying the hanning smoother produces the sequence

4.25, 3.8125, 3.4375, 3.125, 3, 3, 3

The second entry from the left is calculated as:

$$3.8125 = \frac{1}{4}(4.25) + \frac{1}{2}(3.75) + \frac{1}{4}(3.5)$$

6. *Twice* - Smoothers based upon running medians generally tend to cause too much smoothing in a sequence and thereby remove interesting patterns. Recall that the original data set can be envisioned as being decomposed as:

   data = smooth + rough

   To recover patterns from the original series that may be still contained in the rough, one can smooth the rough sequence and then add the result to the smoothed series. Hopefully, key patterns that may have been smoothed away during the first pass of smoothing can be recovered from the rough in this manner. This operation is referred to as *reroughing*.

   For the 4253*H*, twice smoother being considered in this section, the word *twice* indicates the following calculations:

   (i) rough = data − 4253*H*,

   (ii) apply the 4253*H* smoother to the rough, and

   (iii) final smooth = 4253H (applied to given series) + 4253H (applied to rough).

   In the application the 4253*H* smooth is listed in step 5. By comparing this sequence to the original series, the rough values are found to be

   0.75, −1.8125, 0.5625, 0.875, −3, −1, 0, 1

   The first entry on the left, for example, is calculated as:

   $$0.75 = (5 - 4.25).$$

   while the third entry is:

   $$0.5625 = (4 - 3.4375)$$

   Next, the 4253*H* smoother is applied to the rough by using steps 1 to 5 with the rough data. The sequences calculated at each step are listed below:

   Step 1: Apply 4 Smoothing

   0.75, −0.53125, 0.65625, −0.625, −0.21875, −0.5, −0.5,−0.5, 1

   Step 2: Apply 2 Smoothing

0.75, 0.0625, 0.015625, –0.421875, –0.359375, –0.5, 0.1

Step 3: Apply 5 Smoothing

0.75, 0.0625, 0.015625, –0.359375, –0.359375, –0.359375, 0, 1

Step 4: Apply 3 Smoothing

0.15625, 0.0625, 0.015625, –0.359375, –0.359375, –0.359375, 0, 0.71875

Step 5: Apply Hanning

0.15625, 0.07421875, –0.06640625, –0.265625, –0.359375, –0.26953125,

–0.08984375, –0.71875

After applying the 4253*H* smooth to the rough, the last stage is to produce the final smooth by adding this to the 4253*H* smooth of the original data to get:

Final 4253*H*, twice Smooth

4.40625, 3.88671875, 3.37109375, 2.859375, 2.640625, 2.73046875, 3.08984375,

3.71875

Figures 22.3.9 and 22.3.10 display the original hypothetical series and the 4253*H*, twice smooth of the data used in the application. As can be seen, the large amount of rough contained in the given series is eliminated by using the 4253*H* twice smoother.

Electricity Consumption Application - The total annual electricity consumption for the U.S.A. is available from 1920 to 1970 in millions of kiloWatt-hours (United States Bureau of Census, 1976) and a plot of the series is displayed in Figure 4.3.10. As explained in Section 4.3.3, the most appropriate nonstationary model to fit to the square roots of this data set is an ARIMA(0,2,1) model.

The 4253*H*, twice graph of the electrical consumption series without a data transformation is shown in Figure 22.3.11. When compared to the original series depicted in Figure 4.3.10, the smooth is similar in shape to the given highly nonstationary time series. In fact, in both Figures 22.3.11 and 4.3.10, there are clearly visible trends that are increasing dramatically over time. The rough for the 4253*H*, twice smooth can be calculated using

rough = data – 4253*H*, twice.

As shown in Figure 22.3.12, there is indeed a significant rough for the series which is increasing in variance with time. In order to cause the variance to become constant or homoscedastic with time, a square root transformation is required. When comparing Figures 22.3.11 and 22.3.12, the reader should keep in mind that different multiplication factors (10^6 and 10^3, respectively) are used on the ordinate axes. Nonetheless, this example clearly illustrates that benefits can be gained by examining both the smooth and the rough for a given series.

Summary - By following steps 1 to 5, a 4253*H* smooth can be obtained for a given series. To ensure that some key characteristics of the data are not missed, in step 6 the 4253*H* smooth is applied to the rough of the smooth obtained for the original data and then the resulting smooth is added to the first 4253*H* smooth to get the 4253*H*, twice smooth. In

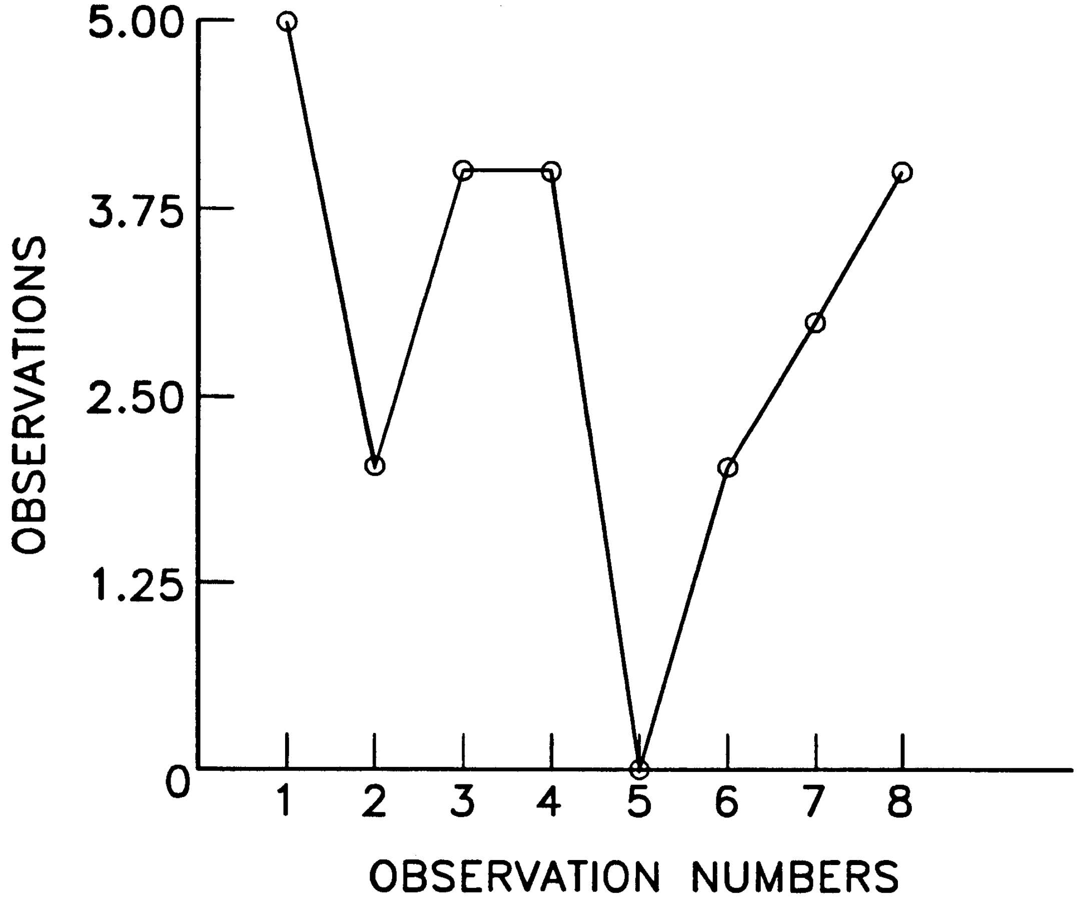

Figure 22.3.9. Graph of the original hypothetical data used in the 4253*H* smoothing example.

addition to examining plots of the 4253*H*, twice smooth to study the main statistical properties of the data, insights can also be gained by studying a plot of the rough for the 4253*H*, twice smooth.

22.3.6 Autocorrelation Function

The *ACF* at lag *k* for a given time series reflects the linear dependence between values which are separated by *k* time lags. The estimate for the ACF at lag *k* for an evenly spaced series, x_t, of length *n* can be calculated using [2.5.9] as (Jenkins and Watts, 1968)

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad k > 0 \qquad [22.3.5]$$

where $\bar{x}$ is the estimated mean of the x_t series. As noted in Section 2.5.4, the value of r_k can range from -1 to +1 where r_0 has a value of unity. Because the ACF is symmetrical about lag

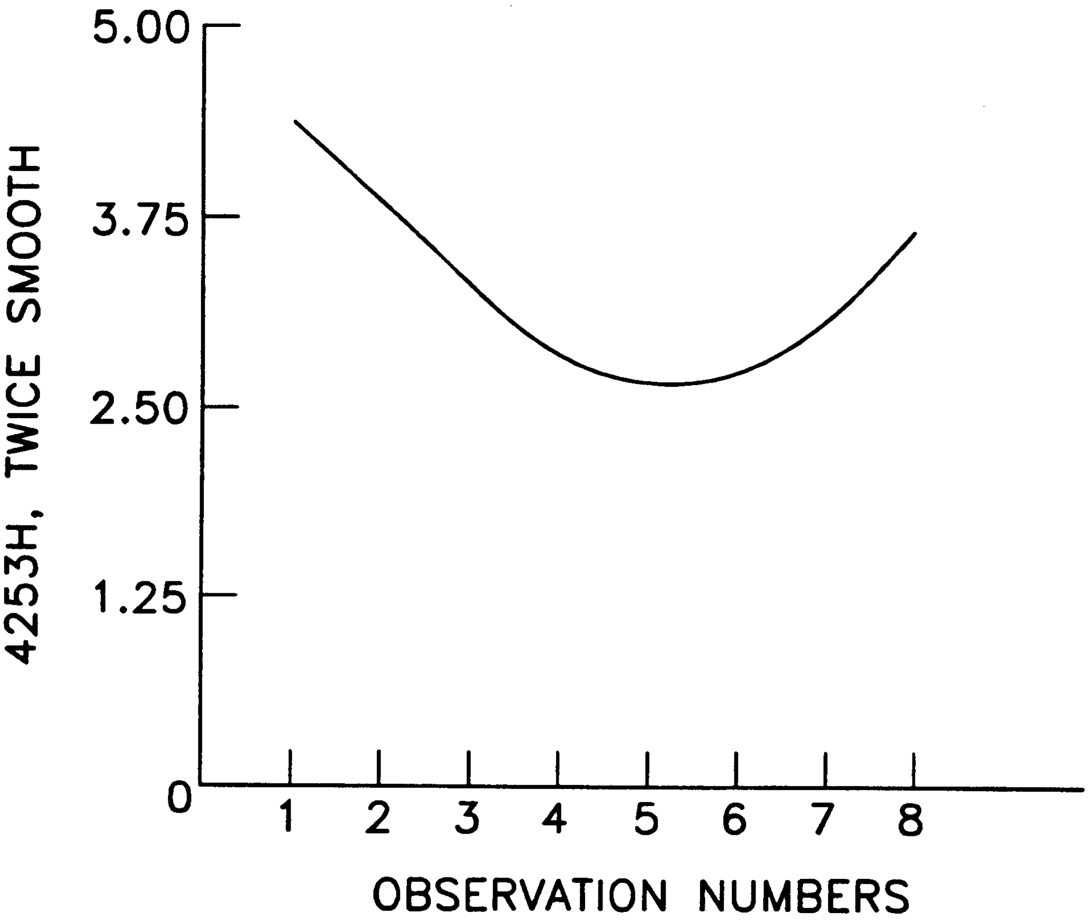

Figure 22.3.10. Plot of the 4253*H*, twice smooth for the hypothetical data of Figure 22.3.9.

zero, it is only plotted for positive lags. When the theoretical ACF is zero and, therefore, the series is white noise, r_k is asymptotically normally independently distributed with a mean of zero and variance of $1/n$. Using simulation experiments, Cox (1966) demonstrated that when r_1 is calculated for a sequence of uncorrelated samples the sampling distribution of r_1 is very stable under changes of distribution and the asymptotic normal form of the sampling distribution is a reasonable approximation even in samples as small as ten.

The ACF furnishes a method for interpreting trends in the data. If, for example, there is a large positive correlation at lag one, this means that in the plot of a series a sequence of high values will often be grouped together and low values will frequently follow other low values. In other words, when r_1 and sample ACF's at other lags are significantly different from zero, this indicates the presence of stochastic trends in the data (see Section 23.1 as well as Section 4.6 for

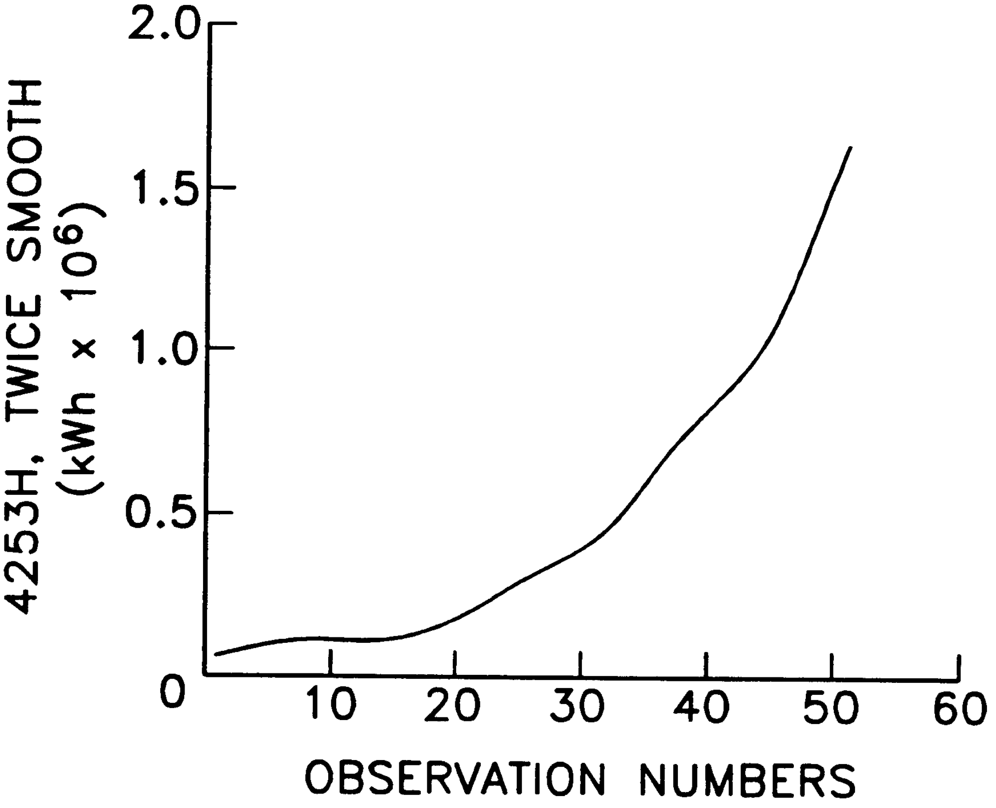

Figure 22.3.11. Plot of the 4253*H* smooth of total annual electricity consumption in the U.S.A. from 1920 to 1970.

discussions of stochastic and deterministic trends). If, for instance, the significance level is less than 0.05 this means that r_1 is significantly different from zero at the 5% significance level. The value of r_1 for the annual total organic carbon series in Figure 20.3.6 is 0.371 with a significance level of 0.137. Consequently, because the significance level of 0.137 is much larger than say the 0.05 significance level, then r_1 is not significantly different from zero. When there is an intervention which causes a significant change in the mean level of a time series such as the change shown in Figures 22.3.6 and 22.3.7 for the total organic carbon series, this introduces a trend in the data due to the observations fluctuating about different mean levels at specified sections in the series. This enforced step trend should cause a rather large value for r_1 for the entire series which is the case for the total organic carbon series. Likewise, an overall trend in the data can cause r_1 to be large. In Section 23.4, simulation experiments demonstrate that the parametric

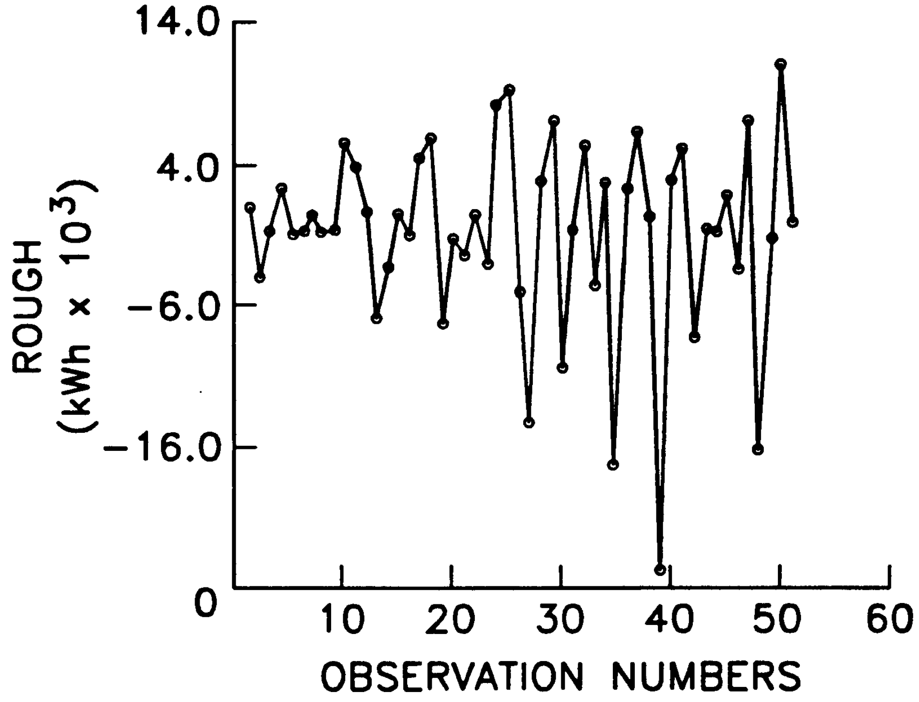

Figure 22.3.12. Rough plot for the 4253H, twice smooth of the total annual electricity consumption in the U.S.A. from 1920 to 1970.

test using r_1 is more powerful than the nonparametric Mann-Kendall test presented in Section 23.3.2 for detecting stochastic trends (see Sections 23.1 and 23.3.1 for a discussion and comparison of parametric and nonparametric tests). However, the Mann-Kendall test is more powerful for discovering deterministic trends.

22.4 CONFIRMATORY DATA ANALYSIS USING INTERVENTION ANALYSIS

22.4.1 Introduction

At the *exploratory data analysis* stage, important statistical characteristics of the data are discovered by employing simple graphical and numerical procedures such as those presented in Section 22.3. Figures 22.3.1 to 22.3.8 show how different exploratory data analysis tools can effectively reveal various statistical properties of the water quality time series which are studied in Section 22.3. Moreover, Figures 23.3.9 to 23.3.12 demonstrate how the $4253H$, twice smooth and also the rough which accompanies this smooth, can uncover interesting statistical characteristics contained in a given time series. When sufficient data are available, confirmatory data analyses can be executed subsequent to the completion of the exploratory data analyses. The main objective of the *confirmatory data analysis* stage is to confirm statistically in a rigorous manner the absence or presence of certain statistical properties in the data which were uncovered by exploratory data analyses. For example, in this section the confirmatory data analysis technique of intervention analysis is used to test the hypothesis that cutting down a forest caused significant changes in the mean levels of certain water quality time series. As noted in Section 22.1, the foregoing systematic approach to data analysis is analogous to a detective solving a crime. At the exploratory data analysis stage, the sleuth collects and studies evidence which is used in rigorous court proceedings at the confirmatory data analysis stage in order to convict the suspected criminal.

Three main approaches to confirmatory data analysis are discussed in Part X of the book. The first technique is *intervention analysis* which is explained in detail in Chapter 19. As noted in Section 19.1, one of the main purposes of intervention analysis is to ascertain whether or not one or more external interventions have caused significant changes in the mean level of a time series. Another use for intervention analysis is to estimate missing observations when there are not a great number of missing values. Due to these and other uses outlined in Section 19.1, intervention analysis is extremely versatile for solving problems in an *environmental impact assessment study*. Within the water quality and quantity applications in Section 22.4.2, intervention analysis is employed for assessing the stochastic effects of an external intervention and also estimating missing values. The different forms of the intervention model used in Section 22.4.2 are described in Sections 19.3 and 19.5 of Chapter 19.

When employing intervention analysis by itself, not more than about 5% of the observations should be missing. If this is the case, intervention analysis constitutes a powerful tool for data filling and also assessing the statistical effects of interventions upon the mean level of a time series. When there are a great number of missing observations, which is the situation for the water quality time series studied in this chapter, the seasonal adjustment algorithm of Section 22.2 can be used to estimate the entries of a time series consisting of equally spaced observations. Subsequent to this, intervention analysis as well as other data analysis tools which can only be used with evenly spaced observations, can be employed.

The second and third major approaches to confirmatory data analysis are the *nonparametric tests* and the *regression analysis* methods described in Chapters 23 and 24, respectively. Although these procedures may not be as powerful as intervention analysis for checking for trends, they do not require the observations to be evenly spaced over time. Consequently, the nonparametric tests and regression analysis can be used with either evenly or unevenly spaced observations and no data filling is required.

22.4.2 Intervention Analysis Applications

Case Study

In 1961 the Marmot Creek experimental basin was established on the eastern slopes of the Rocky Mountains in Alberta, Canada (Jeffrey, 1965; Golding, 1980). The objective of the study was to determine the hydrology of the area so that guidelines which are consistent with the importance of the eastern slopes as a water supply area for Alberta and Saskatchewan, could be formulated for harvesting trees. Both the Middle Fork and Cabin Creeks are located within the Marmot basin in the province of Alberta and flows in these creeks are unregulated. Upstream from the gauging station, the area of the forested Middle Fork basin is 2.85 km^2 while the upstream area of the Cabin Creek basin is 2.12 km^2. From July to October, 1974, an intervention took place in the Cabin Creek basin when 40% of the forested area was clear-cut. Because the trees in the forested Middle Fork basin were not cut down and the basin is located close to the Cabin Creek basin, the appropriate water quality and quantity series from the Middle Fork Creek can be used as covariate series for intervention models developed for the Cabin Creek data sets. In this way the intervention components in the intervention models will more accurately measure the effects of the intervention in the Cabin Creek series.

Three different types of intervention models are developed in this section and also by McLeod et al. (1983) for solving various aspects of the problem created in the Marmot basin due to cutting down the trees in the Cabin Creek basin. The most important and interesting of the three intervention models is the third one which is called the *General Water Quality Intervention Model*. For a given water quality series for the Cabin Creek, the purpose of the third intervention model is to rigorously ascertain for which months the forest cutting intervention caused significant changes in the mean level of the water quality series. As will be seen, in order to calibrate this model, complete average monthly flow records are needed for the Cabin Creek flows which in turn are closely correlated with the Middle Fork Flows. Because there are eight missing values for the average monthly flows of Middle Fork River, an intervention model similar to the one in Section 19.3 is constructed in order to obtain efficient estimates for the missing observations. Following the development of the *Middle Fork Flow Intervention Model, The Cabin Creek Flow Intervention Model* is built for estimating four missing values in the time series of average monthly flows of the Cabin Creek and also for determining the effects of the clear-cutting intervention upon the Cabin Creek flows. In order to increase the accuracy of the Cabin Creek Flow Intervention Model, the average monthly flows of the Middle Fork River are used as a covariate series. Hence, this model is similar to the one described in detail in Section 19.5. Subsequent to the completion of the first two water quantity intervention models, a *General Water Quality Intervention Model* can be built for each of the water quality variables measured in the Cabin Creek. Because each of the water quality series consists of observations which are unevenly spaced, the data filling technique of Section 22.2 can be utilized for estimating a sequence of average monthly values. Water quality models are developed for all of the time series except the dissolved iron series, since no data are available after the intervention for this series, and also the extractable iron series. For each water quality intervention model, the covariate series are the same water quality series for the Middle Fork basin and also the monthly flows of the Cabin Creek. As an illustrative example, the procedure for fitting an intervention model to the total organic carbon series on the Cabin River, is fully explained. This water quality model constitutes an interesting version of the types of intervention models presented in Section 19.5.

Middle Fork Flow Intervention Model

Before the Middle Fork flows can be used as a covariate series in an intervention model for the Cabin Creek flows, the missing observations for the Middle Fork Creek must be estimated using a separate intervention model similar to the one in [19.3.5]. Average monthly flows are unknown for the Middle Fork Creek for April and May of 1974, February, March and April of

1975, and for January, February, and March of 1978. The entire data set for the Middle Fork flows extends from January 1964 to December 1979.

Following the notation used in Chapter 19, let y_t represent the logarithmic monthly flow of the Middle Fork Creek at time t. From [19.3.5], an intervention model for estimating the unknown observations has the form

$$y_t - \bar{y} = \sum_{j=1}^{8} \omega_{0j} \xi_{tj} + N_t \qquad [22.4.1]$$

where $\bar{y}$ is the mean of the entire y_t series, ω_{0j} is the parameter of the jth transfer function, ξ_{tj} is the jth intervention series which is assigned a value of unity where the jth observation is missing and zero elsewhere, and N_t is the noise component.

To identify the noise term, N_t, in [22.4.1], a seasonal ARIMA or SARIMA model can be fitted to the series prior to the first missing value, by utilizing appropriate model construction tools from Section 22.3. When fitting a SARIMA model to average monthly riverflows, usually it is necessary to take natural logarithms of the data and then to difference the transformed series seasonally using the operator defined in [12.2.3]. Figure 22.4.1 is a plot of the sample ACF for the seasonally differenced logarithmic data for the period from January, 1964 to December, 1973. The ACF can be calculated by substituting the data into [2.5.9] and the 95% confidence limits in Figure 22.4.1 are calculated using [12.3.1] under the assumption that the theoretical ACF is zero after lag $k = 13$. Because the ACF attenuates starting at lag 1, this may indicate the need for a nonseasonal AR parameter. Since there is a large value of the sample ACF at lag 12, this indicates that a seasonal MA parameter may be required. Consequently, following the notation from Section 12.2, it may be appropriate to fit a SARIMA $(1,0,0)\times(0,1,1)_{12}$ model from [12.2.7] to the series. The sample PACF presented in Sections 3.2.2 and 12.3.2, also confirms that this may be a reasonable model. After fitting the SARIMA $(1,0,0)\times(0,1,1)_{12}$ model to the Middle Fork data before the intervention, the ACF for the residuals can be calculated. In Figure 22.4.2, the 95% confidence limits are calculated using the formula given in [12.3.7]. Except for the residual ACF value at lag 24, all of the values lie within the 95% confidence limits and, therefore, the assumption of white noise is satisfied. The slightly large value at lag 24 could be due to chance.

Following the identification of N_t for use in [22.4.1], all of the parameters in the intervention model can be simultaneously estimated. Because natural logarithms are taken of the data, a positive quantity must be placed at the location where the observations are missing. For convenience, the logarithm of the appropriate monthly mean is substituted for each missing value in the series. In the second column of Table 22.4.1, the MLE's (maximum likelihood estimates) of the parameters and SE's (standard errors) are given for the eight ω_{0j} intervention parameters in [22.4.1]. From Section 9.3.3, the estimate of the missing observation in the logarithmic or transformed domain is

$$\hat{y}_{t_j} = \ln \textit{monthly mean} + \hat{\omega}_{0j}$$

where t_j is the time for which the observation at t_j is missing. By taking the inverse logarithmic transformation, the estimate of the missing observation in the untransformed domain is

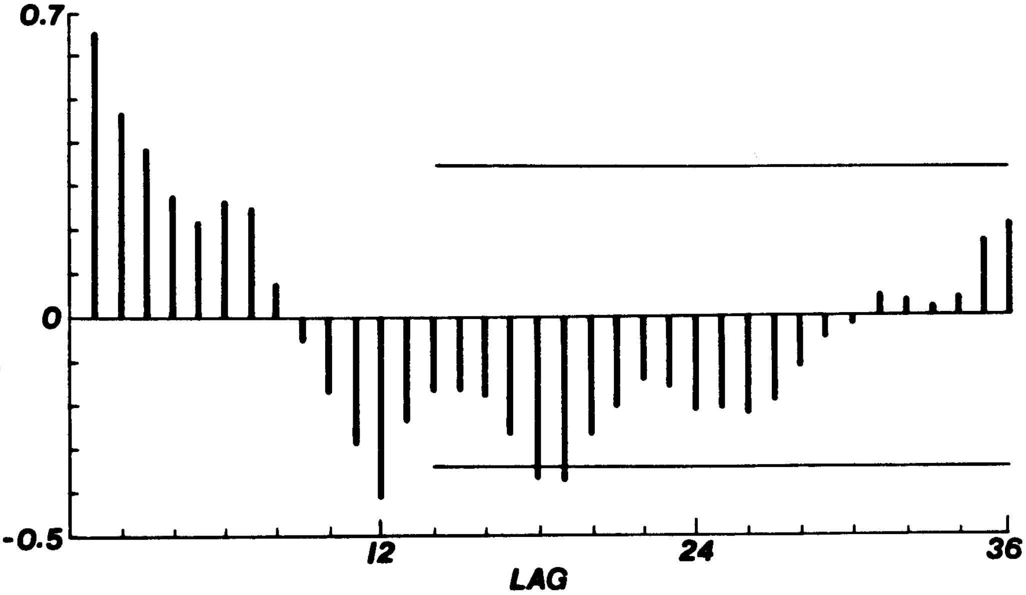

Figure 22.4.1. Sample ACF for the seasonally differenced logarithmic Middle Fork flows before the intervention.

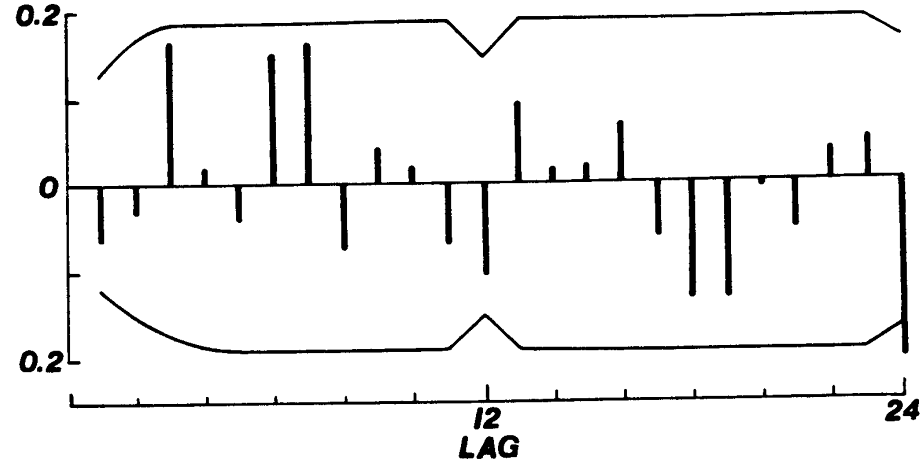

Figure 22.4.2. Residual ACF for the Middle Fork SARIMA model for before the intervention.

$$\hat{Y}_{t_j} = e^{\hat{y}_{t_j}} = \exp(\ln \textit{monthly mean} + \hat{\omega}_{0j}) \qquad [22.4.2]$$

The estimates for the missing monthly observations in m^3/s are displayed in the last column in Table 22.4.1.

Table 22.4.1. Estimated parameters for the Middle Fork Flow intervention model.

Parameter	Estimate of ω_{0j} ± Standard Error	Date of Missing Value	Estimate of Missing Value in m^3/s
ω_{01}	0.6029 ± 0.2976	April, 1974	0.0219
ω_{02}	-0.5316 ± 0.2959	May, 1974	0.0348
ω_{03}	0.1849 ± 0.3083	Feb., 1975	0.0048
ω_{04}	0.3057 ± 0.3456	March, 1975	0.0041
ω_{05}	0.6885 ± 0.3084	April, 1975	0.0139
ω_{06}	-0.1244 ± 0.3150	Jan., 1978	0.0044
ω_{07}	-0.0236 ± 0.3571	Feb., 1978	0.0039
ω_{08}	-0.0838 ± 0.3164	March, 1978	0.0028

Cabin Creek Flow Intervention Model

Because the missing values in the Middle Fork riverflow series have all been estimated, the complete data set can now be used as a covariate series for an intervention model for the Cabin Creek. However, there are four missing observations for the Cabin Creek which occur in February, March, April, and May of 1979. Consequently, in addition to the clear-cutting intervention which took place from July to October of 1974, intervention components must be included in the model so that the four missing values can be estimated.

Some of the exploratory data analysis tools from Section 22.3 can be employed to check if there appear to be changes in the Cabin Creek flows due to the forest cutting intervention. For example, box-and-whisker graphs from Section 22.3.3 can be constructed for the series both before and after the intervention date. When the medians for each month before and after the intervention are compared, the median levels do not appear to change very much. Likewise, when other exploratory tools are employed, no noticeable changes in the flow series are detected as a result of cutting down the forest.

To ascertain rigorously if clear-cutting the forest has significantly affected the mean levels of the average monthly flows of the Cabin Creek, an appropriate intervention model can be constructed at the confirmatory data analysis stage by following the three steps of identification, estimation and diagnostic checking described in Section 19.5.3. Following the general format of the model in [19.5.8], the intervention model for the flows of the Cabin Creek can be written as

$$y_t - \bar{y} = \sum_{i=1}^{12} \omega_{0i} \xi_{ti} + \sum_{j=13}^{16} \omega_{0j} \xi_{tj} + \omega_{017}(x_t - \bar{x}) + N_t \qquad [22.4.3]$$

where y_t is the monthly logarithmic Cabin Creek flows and $\bar{y}$ is the mean of the entire y_t series, ω_{0i} is the transfer function parameter for the ith monthly intervention where there is one parameter for each month of the year, ξ_{ti} is a monthly step intervention which is given a value of unity for the month it represents during and after the intervention but given a value of zero elsewhere, ω_{0j} is the transfer function parameter for a missing data point where there are four such parameters, ξ_{tj} is the intervention series for ω_{0j} where it is given a value of unity at the time that the observation is missing and a value of zero elsewhere, x_t is the logarithmic Middle Fork flows and $\bar{x}$ is the mean of the entire x_t series, ω_{017} is the transfer function parameter for the covariate x_t series, and N_t is the correlated noise term.

The transfer functions in [22.4.3] are designed from a physical understanding of the problem. Since it would be expected that the Middle Fork and Cabin Creek flows would be closely related during the same month due to common climatic conditions, the parameter ω_{017} is included as the parameter in the covariate transfer function. Because only six observations for each month are available after the intervention, the step intervention series, ξ_{ti}, along with the ω_{0i} parameter is included in the first term for each month on the right hand side in [22.4.3]. When more data becomes available, it may be reasonable to include a parameter in the denominator of each transfer function that models the clear-cutting intervention. In [19.5.10], it is shown how a term in the denominator can model the attenuating affects of a forest fire upon riverflows as the forest slowly recovers over the years. In this study, transfer functions of the form $\dfrac{\omega_{0i}}{1 - \delta_{ii} B^{12}}$ were included in the first summation term on the right hand side of [22.4.3] but meaningful results were not obtained due to the lack of sufficient data and a long enough time period after the intervention. Perhaps after about ten years, enough data will be available so that two parameters can be included and thereby allow the impacts of forest recovery to be more fully explored within the structure of the intervention model.

After designing the transfer functions in [22.4.3], the noise term, N_t, must be identified. As noted throughout Chapter 19, a convenient procedure to employ is to first assume that N_t is white noise. The parameters and residuals in [22.4.3] can then be estimated. Since the residuals will probably not be white noise, a SARIMA model can be identified for fitting to the residuals. For the case of the Cabin Creek residuals, the most appropriate model is found to be a SARIMA $(1,0,0)\times(1,0,0)_{12}$ model. Assuming this form for N_t, the parameters for the complete model in [22.4.3] are simultaneously estimated again. Diagnostic checks applied to the residuals from the latest model design, demonstrate that the model is satisfactory since the residuals are uncorrelated. In practice, various forms of the transfer functions and noise term must be tried before a satisfactory model is found. Consequently, the model building procedure is not quite as simple as it may appear in the foregoing explanation. Experience, coupled with a sound understanding of both the physical problem and the capabilities of the intervention model, help to reduce the time required to design an appropriate intervention model.

The parameter estimates and SE's for the four missing observations are given in Table 22.4.2. Since natural logarithms are taken of the data, it is necessary to include positive values at the four locations where the observations are missing. The logarithm of the appropriate average monthly value across all the years is substituted at each location where an observation is missing. After calibrating the complete intervention model written in [22.4.3], each estimated missing value can be calculated using [22.4.2]. In the last column in Table 22.4.2, the estimated monthly values ore displayed.

Table 22.4.2. Estimated missing values for the Cabin Creek flows.

Parameter	Estimate ± Standard Error	Date of Missing Value	Estimate of Missing Value
ω_{013}	0.2893 ± 0.2300	Feb., 1979	0.004
ω_{014}	0.1351 ± 0.2607	March, 1979	0.003
ω_{015}	0.6682 ± 0.2611	April, 1979	0.012
ω_{016}	0.5834 ± 0.2300	May, 1979	0.079

The MLE for ω_{017}, which is the transfer function parameter for the covariate series consisting of the logarithmic Middle Fork flows, is 0.814 with a SE of 0.020. Because $\hat{\omega}_{017}$ is much larger than 1.96 times the SE, it is significantly different from zero. Accordingly, it is worthwhile to include the covariate series in the intervention model in [22.4.3] in order to enhance the credibility and accuracy of the model.

The MLE's and SE's for the twelve intervention parameters for the clear-cutting, are presented in Table 22.4.3. To calculate the percentage change in the mean level of a specific monthly flow due to the intervention, the following formula given in [19.2.20] is employed.

$$\% \; change = (e^{\hat{\omega}_{0i}} - 1)100 \qquad [22.4.4]$$

To calculate the 95% confidence limits simply add and subtract 1.96 times the SE to the estimated ω_{0i} and then substitute these two values into [22.4.4]. The percentage change in the mean level for each month along with the 95% confidence levels are presented in Table 22.4.3. For nine out of twelve months, zero falls within the 95% confidence limits. Therefore, for these months it can be argued that the percentage changes in the mean levels are not significantly different from zero. However, for January, March and November, zero does not fall within the 95% confidence limits and, consequently, there appear to be significant changes in the mean levels for these months. For January and March the mean levels have decreased while for November the average flow has risen. Nevertheless, notice that for each of these three months, one side of the limits for the 95% confidence limits, is quite close to zero. Consequently, to simplify the intervention model developed in the next subsection, it is assumed that the clear cutting of the forest has not significantly altered the Cabin Creek flows.

Table 22.4.3. Estimated parameters for modelling the intervention effects in the Cabin Creek flow intervention model.

Month	Parameter	Estimate ± Standard Error	Percentage Change	95% Confidence Interval
Jan.	ω_{01}	-0.2303 ± 0.1153	-20.57	-36.64, -0.43
Feb.	ω_{02}	-0.1930 ± 0.1183	-17.55	-34.62, 3.97
March	ω_{03}	-0.2353 ± 0.1191	-20.96	-37.42, -0.18
April	ω_{04}	-0.1652 ± 0.1192	-15.23	-32.88, 7.07
May	ω_{05}	-0.0931 ± 0.1197	-8.89	-27.94, 15.20
June	ω_{06}	-0.0682 ± 0.1206	-6.59	-26.25, 18.31
July	ω_{07}	-0.0872 ± 0.1290	-8.35	-28.83, 18.01
Aug.	ω_{08}	-0.1353 ± 0.1411	-12.65	-33.75, 15.17
Sep.	ω_{09}	0.0802 ± 0.1454	8.35	-18.51, 44.08
Oct.	ω_{010}	-0.1468 ± 0.1419	-13.65	-34.62, 14.04
Nov.	ω_{011}	0.3007 ± 0.1396	35.08	2.75, 77.58
Dec.	ω_{012}	-0.0337 ± 0.1289	-3.31	-24.90, 24.49

General Water Quality Intervention Model

Intervention models were developed for twelve water quality variables on the Cabin Creek although representative results are only shown in this section for the total organic carbon intervention model. For each water quality intervention model, the covariate series are the same water quality series for the Middle Fork basin and also the monthly flows of the Cabin Creek. Qualitatively, the General Water Quality Intervention Model is written as

$$\begin{array}{lcccccccc}
\textit{Cabin} & & & & & & \textit{Middle} & & \\
\textit{Creek} & & \textit{monthly} & & \textit{Cabin} & & \textit{Fork} & & \\
\textit{water} & = & \textit{interventions} & + & \textit{Creek} & + & \textit{water} & + & \textit{Noise} \\
\textit{quality} & & & & \textit{flows} & & \textit{quality} & & \\
\textit{series} & & & & & & \textit{series} & &
\end{array}$$

Mathematically, appropriate components from the finite difference equation in [19.5.2] can be utilized in order to write the General Water Quality Intervention Model as

$$y_t - \bar{y} = \sum_{i=1}^{12} \omega_{0i}\xi_{ti} + \omega_{013}(x_{t1} - \bar{x}_1) + \omega_{014}(x_{t2} - \bar{x}_2) + N_t \qquad [22.4.5]$$

where y_t is the average monthly water quality series for the Cabin Creek that was estimated using the seasonal adjustment algorithm of Section 22.2, $\bar{y}$ is the mean of the y_t series, ξ_{ti} is the intervention series for a given month where it is assigned a value of one for the month it represents from the intervention onwards and a value of zero elsewhere, ω_{0i} is the transfer function parameter for the ξ_{ti} series and the MLE for ω_{0i} can be used to ascertain the effects of the intervention for the month being studied, x_{t1} is the estimated monthly logarithmic series for the

Cabin Creek where the seasonal adjustment algorithm in Section 22.2 is used to estimate the monthly flows from daily flows that occur at the same time as the water quality observations, $\bar{x}_1$ is the mean of the x_{t1} series, ω_{013} is the transfer function for the Cabin Creek flow series, x_{t2} is the same estimated monthly water quality series as y_t but for the Middle Fork Creek and the seasonal adjustment algorithm is used to estimate x_{t2}, $\bar{x}_2$ is the mean of the x_{t2} series, ω_{014} is the transfer function parameter for the covariate Middle Fork water quality series, and N_t is the noise term which can be modelled by an appropriate SARMA or SARIMA model from Chapter 12.

In [22.4.5] the seasonally adjusted monthly flows are employed as a covariate series, x_{ti}. The reason for using the seasonally adjusted series rather than the known monthly riverflows is that this may help to eliminate any problems due to seasonal adjustment that are contained in the y_t series. It should be kept in mind that by considering the flows as a covariate series, the stochastic or statistical relationship between the flow, x_{t1}, and the water quality series, y_t, is formally modelled through the transfer function parameter, ω_{013}, in the overall intervention model in [22.4.5].

When constructing the water quality intervention models in [22.4.5], the identification, estimation and diagnostic check stages of model development described in Section 19.5.3 are adhered to. Although the transfer functions for all the water quality series are the same as those in [22.4.5], it should be pointed out that quite a few different types of transfer functions were actually tested. For instance, because not too many observations for each month are available after the intervention, a step intervention along with a ω_{0i} parameter is included in the first term for each month on the right hand side in [22.4.5]. As also noted for the intervention model in [22.4.3], if more data were available the possibility of including a parameter in the denominator of each transfer function would have been feasible. In [19.5.10] within Section 19.5.4, it is explained how a term in the denominator can model the attenuating effects of a forest fire upon riverflows as the forest slowly recovers over the years. Finally, a specific SARIMA model had to be identified separately for modelling N_t in [22.4.5] for each water quality intervention model. The same procedure described for the Cabin Creek flow intervention model was employed to design a specific noise term for each intervention model.

Total Organic Carbon Application: As shown by the applications in Section 22.3, exploratory data analyses clearly detect the effects of the forest clearing upon the total organic carbon series for the Cabin Creek. For example, when the box-and-whisker graphs for before and after the intervention are compared in Figures 22.3.4 and 22.3.5, respectively, the decrease in the median level after the intervention can be easily seen for almost all the months. Likewise, the average annual plot in Figure 22.3.6 and the blurred smooth in Figure 22.3.7 clearly detect the drop in the mean level of total organic carbon in later years. Finally, since the value of the ACF at lag one calculated using [22.3.5] for the annual series is significantly different from zero, this suggests the presence of a trend in the data.

The foregoing exploratory facts are rigorously confirmed in a statistical sense by fitting the intervention model in [22.4.5] to the total organic carbon series which is available from the start of 1971 to the end of 1978. Natural logarithms are used for the two total organic carbon series given by y_t and x_{t2} for the Cabin and Middle Fork Creeks, respectively. The SARIMA model identified for the noise term, N_t, contains one nonseasonal AR parameter and one seasonal AR parameter, and as explained in Section 12.2.2, it can be written as $(1,0,0)\times(1,0,0)_{12}$. The parameter, ω_{013}, which relates the Cabin Creek flows to the total organic carbon in the Cabin Creek has a MLE of 0.081 with a SE of 0.095. Since the MLE of ω_{013} is about the same size as its SE, it may be worthwhile to include the flows as a covariate series in the intervention model. The MLE for ω_{014} is 0.620 with a SE of 0.082 and, consequently, it is very informative to incorporate the covariate total organic series from Middle Fork Creek into the model. In Table 23.4.4, the MLE's and SE's are presented for the twelve intervention parameters contained in the first component on the right hand side of [22.4.5]. Also included in Table 23.4.4 is the percentage change in mean level for each month along with the 95% confidence limits which are calculated using [22.4.4]. For all the months where zero is not included in the 95% confidence limits, the percentage change in the mean level is confirmed to be significantly different from zero. Accordingly, from Table 22.4.4 it can be seen that there is a significant drop in the mean level of total organic carbon in the Cabin Creek during the summer months of June, July and August.

Table 22.4.4. Intervention parameter estimates for the total organic carbon intervention model for the Cabin Creek.

Month	Parameter	MLE	Standard Error	Percentage Change	95% Confidence Limits	
January	ω_{01}	0.002	0.231	0.17	-36.33,	57.59
February	ω_{02}	0.085	0.231	8.92	-30.71,	71.22
March	ω_{03}	-0.169	0.227	-15.52	-45.86,	31.82
April	ω_{04}	-0.216	0.224	-19.42	-48.11,	25.11
May	ω_{05}	-0.053	0.228	-5.12	-39.25,	48.20
June	ω_{06}	-0.716	0.227	-51.12	-68.69,	-23.68
July	ω_{07}	-0.524	0.256	-40.81	-64.18,	-2.20
August	ω_{08}	-0.566	0.260	-43.21	-65.88,	-5.49
September	ω_{09}	0.026	0.260	2.65	-38.34,	70.90
October	ω_{010}	-0.019	0.258	-1.91	-40.86,	62.69
November	ω_{011}	0.048	0.266	4.90	-37.75,	76.79
December	ω_{012}	-0.344	0.270	-29.13	-58.26,	20.32

22.5 CONCLUSIONS

To execute a comprehensive data analysis study, one can follow the exploratory data analysis and confirmatory data analysis stages. As demonstrated by water quality applications, this approach is especially effective for detecting and modelling trends which may be contained in messy environmental data. The time series being analyzed may be very messy because the series may possess various handicaps such as having missing observations, being nonnormally distributed, possessing outliers and being short in length. Nevertheless, by employing appropriate exploratory and confirmatory data analysis tools, as much useful information as possible can be gleaned from the available data, even if the quality and quantity of the data are not very good.

The purpose of the exploratory data analysis stage is to uncover important statistical properties of the data by utilizing simple graphical and numerical tools, some of which are discussed in detail in Section 22.3. Some exploratory techniques such as a graph of the series against time (Section 22.3.2), box-and-whisker graphs (Section 22.3.3) and the cross-correlation function (Section 22.3.4), do not require that the time series be evenly spaced over time. On the other hand, other exploratory data analysis techniques like Tukey smoothing (Section 22.3.5) and the ACF (Section 22.3.6) are designed to be used with data points that are equally spaced over time. Fortunately, a number of flexible data filling procedures are now available for estimating the entries of an evenly spaced time series from a data set for which the time intervals between adjacent observations are not the same. Depending upon how much information is missing and the number of observations available, an appropriate data filling technique can be selected from Section 22.2, 19.3, or 18.5.2. Subsequent to filling in missing observations, one can employ suitable exploratory and confirmatory data analysis tools which require evenly spaced measurements.

At the confirmatory data analysis stage, three different types of approaches which can be used to rigorously characterize trends are intervention analysis, nonparametric tests and regression analysis, described in Section 22.4 and Chapter 19, Chapter 23, and Chapter 24, respectively. Nonparametric tests and regression analysis can be used with unequally or equally spaced data whereas intervention analysis must be employed with an evenly spaced sequence of observations. As demonstrated in this chapter and also Chapter 19, the intervention model constitutes an extremely powerful and comprehensive parametric model which can accurately model the magnitude and shape of a trend caused by a known intervention. Furthermore, as explained in Section 22.4.2, the impacts of water quantity upon water quality can be realistically incorporated into the intervention model by including the water quantity time series as a covariate series in the intervention model in [22.4.5].

In many situations, an analyst is requested to execute a comprehensive data analysis study in order to discover and model trends after the data have already been collected by other people. As a result, the data may be rather messy and thereby difficult to model. Of course, there are no data analysis tools that can extract information which is not contained in the data to begin with. Nonetheless, by using the most suitable data analysis techniques, the maximum amount of useful information can be discovered and modelled. When the analyst can assist in optimally designing the data collection scheme, then some of the suggestions given in Section 19.7 and elsewhere may be helpful.

PROBLEMS

22.1 The seasonal adjustment algorithm in Section 22.2 is described for estimating average monthly values for a time series using daily values that are available at irregular time intervals. Explain how this algorithm would work for the following situations:

(a) estimating average quarterly values, and

(b) estimating weekly values.

22.2 Select a daily time series for which all of the observations are available over a ten year time period. Randomly remove about 70% of the daily data and then employ the seasonal adjustment algorithm of Section 22.2 to estimate the average monthly values. Compare these monthly estimates to those obtained when the complete set of daily values are employed for determining the average monthly values.

22.3 In Section 22.3, some useful exploratory data analysis tools are presented. By referring to an appropriate reference on exploratory data analysis, describe three other exploratory data analysis techniques which are not discussed in Section 22.3. You may, for instance, wish to write about stem-and-leaf displays. Be sure to mention the main statistical characteristics that each method is designed to uncover in a given data set.

22.4 Select an average monthly time series that is of interest to you. Obtain a plot of the observations over time as well as a box-and-whisker graph for each season. Describe the main statistical characteristics contained in your data set which are graphically revealed using each exploratory data analysis technique. Which of the two graphical methods was most helpful for better understanding the statistical and stochastic properties of your data?

22.5 As mentioned in Section 22.3.3, a seasonal notched box-and-whisker graph can be employed for graphically testing whether medians across two or more seasons are significantly different. Explain why using seasonal notched box-and-whisker plots in this way is equivalent to graphically carrying out a formal hypothesis test (see Section 23.2.2 for a review of hypothesis tests). In your explanation, be sure to clearly state the null and alternative hypotheses as well as the test statistic. Assuming normality and independence of the data for each season, derive the 5% significance level for the test. Finally, state advantages of graphically implementing statistical tests as part of exploratory data analysis tools.

22.6 Select a set of water quality time series measurements that are available for a variety of water quality variables measured in a river or lake. Following the procedure of Section 22.3.4, determine the cross-correlation matrix for these series. When commenting upon your results use physical explanations of the phenomena to help confirm what is found statistically.

22.7 Choose an annual time series which you suspect may contain trends. For this series, plot the following graphs and then comment upon your findings regarding the main statistical characteristics of the series. Be sure to make comparisons across the graphs and clearly point out the advantages of studying each type of graphical output.

(a) Plot of the data,

(b) Box-and-whisker graph,

(c) Blurred 3RSR smooth,

(d) 4253H, twice smooth,

(e) Rough for the 4253H, twice smooth.

22.8 Carry out the instructions of 22.7 for a seasonal time series of your choice.

22.9 Select an annual time series to which you apply all of the most appropriate exploratory data analysis tools of Section 23.3 and elsewhere. Justify the reasons for choosing these exploratory techniques and summarize your main statistical findings.

22.10 Execute the instructions of problem 22.9 for a seasonal time series.

22.11 By referring to an appropriate reference, find a smoother not covered in Section 22.3.5 which you think may work well in practice. Outline the steps that are followed when applying this smoother to a time series. Assess the main capabilities and weaknesses of the smoother. Finally, apply this smoother to a time series of your choice and comment upon your statistical findings.

22.12 In a column on the left hand side of a page, write down a fairly extensive list of statistical characteristics which you would like exploratory data analysis tools to discover when examining water quality time series. In a row across the top of the page copy down the names of a variety of informative exploratory data analysis techniques. Then, below each technique put check marks opposite the statistical properties that the method is designed to find when these properties are present in the data. Explain how this table that summarizes the capabilities of exploratory data methods can be useful in a case study.

22.13 Carry out the instructions of problem 22.12 for hydrological time series.

22.14 Execute the instructions of problem for 22.12 for meteorological data sets.

22.15 Select a set of water quantity and quality time series that may have been significantly influenced by a known external intervention.. Carry out a comprehensive data analysis of these time series by employing appropriate exploratory and confirmatory data analysis tools in order to detect and model possible trends as well as other interesting statistical characteristics.

REFERENCES

DATA SET

United States Bureau of the Census (1976). *The Statistical History of the United States from Colonial Times to the Present.* United States Government.

EXPERIMENTAL BASINS

Golding, D. L. (1980). Calibration methods for detecting changes in streamflow quantity and regime. In *The Influence of Man on the Hydrological Regime with Special Reference to Representative and Experimental Basins, Proceedings of the Helsinki Symposium,* IAHS (International Association of Hydrological Sciences), 130:3-7.

Jeffrey, W. W. (1965). Experimental water sheds in the Rocky Mountains, Alberta, Canada. In *Proceedings of the Symposium on Representative and Experimental Areas, Budapest Symposium*, 6:502-521 IAHS (International Association of Hydrological Sciences) 66.

EXPLORATORY DATA ANALYSIS

Barnett, V. (1975). Probability plotting methods and order statistics. *Applied Statistics* 24(1):95-108.

Berthouex, P. M., Hunter, W. G., and Pallesen, L. (1981). Wastewater treatment: A review of statistical applications. In *Environmetrics 81: Selected Papers, Selections from USEPA-SIAM-SIMS Conference*, Alexandria, Virginia, pages 77-99.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829-836.

Cluis, D. A. (1983). Visual techniques for the detection of water quality trends: Double-mass curves and CUSUM functions. *Environmental Monitoring and Assessment*, 3:173-184.

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

Cunnane, C. (1978). Unbiased plotting positions - a review. *Journal of Hydrology*, 37:205-222.

du Toit, S. H. C., Steyn, A. G. W., and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis*. Springer-Verlag, New York.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378-385.

Hewlett-Packard (1977). *HP-29C Applications Book*.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

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

Mallows, C. L. (1980). Resistant smoothing. In Anderson, O. D., editor, *Time Series, Proceedings of the International Conference held at Nottingham University*, March 1979, pages 147-155. North-Holland.

McGill, R., Tukey, J. W. and Laren, W. A. (1978). Variation of Box plots. *The American Statistician*, 32(1):12-16.

McNeil, D. R. (1977). *Interactive Data Analysis*. Wiley, New York.

Ramsey, F. L. (1988). The slug trace. *The American Statistician*, 42(4):290.

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

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, 75:609-615.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston.

SEASONAL ADJUSTMENT

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3-32.

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

Hillmer, S. C. and Tiao, G. C. (1985). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association*, 77:63-70.

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

Shiskin, J., Young, A. H. and Musgrave, J. C. (1976). The x-11 variant of the census method II seasonal adjustment program. Technical report number BEA-76-01, Bureau of Economic Analysis, U.S. Department of Commerce, Washington, D.C.

TREND ASSESSMENT STUDIES

McLeod, A. I., Hipel, K. W. and Camacho, F. (1983). Trend assessment of water quality time series. *Water Resources Bulletin*, 19(4):537-547.