CHAPTER 24

REGRESSION ANALYSIS

AND

TREND ASSESSMENT

24.1 INTRODUCTION

Suppose that one would like to analyze trends in water quality time series measured in rivers. The underlying conceptual model used in *trend analysis* of a water quality variable Y can be written as

$$Y_t^{(\lambda)} = f(X_t) + S_t + C_t + \varepsilon_t \qquad [24.1.1]$$

where Y_t is the water quality observation at time t that may be transformed using the Box-Cox transformation in [3.4.30] to produce the transformed value given as $Y_t^{(\lambda)}$, $f(X_t)$ is a function of a covariate series X_t, X_t is a covariate series at time t such as riverflow or temperature measured at time t, S_t is the seasonal component at time t, C_t is the trend in Y_t, and ε_t is the noise component at time t. In trend analysis, one wishes to appropriately account for X_t, S_t and ε_t so that C_t can be easily detected and accurately quantified, even when the trend effects are small. Trends over time can be increasing, decreasing or non-existent. Furthermore, trends over time can follow linear or nonlinear geometrical patterns. For many water quality series measured in rivers, the covariate series used is flow. However, in other situations, different covariate series can be used. For example, temperature may be better to use as X_t when the Y_t series is dissolved oxygen or total nitrates. Also, when water quality measurement sites are far removed from flow gauging sites, covariates other than flow may have to be used for X_t. Finally, the idea of decomposing a series into its basic components as in [24.1.1] is a well established procedure and is, for example, inherent in the basic designs of the intervention models in [19.5.8] and [22.4.5] as well as the seasonal adjustment procedure of Section 22.2.

The objectives of this chapter are twofold. The first goal is to explain some ways in which *regression analysis* can be used to model various components of the general model in [24.1.1]. The second objective is to develop a general *trend analysis methodology* based on [24.1.1] for analyzing trends in water quality time series measured in rivers. As is explained in Section 24.3.2, regression analysis as well as nonparametric tests play a key role in this methodology. To demonstrate the efficacy of the methodology, it is applied to representative water quality time series measured in rivers in Southern Ontario, Canada.

As pointed out in the preface to Part X as well as Section 23.1, water quality and other kinds of environmental data are often quite *messy*. The time series may, for example, be highly skewed, have many missing observations, possess multiple censoring levels, and contain seasonal trends. When examining environmental data for the presence of trends and other statistical properties, a systems design approach to data analysis should be followed. More specifically,

one may wish to adhere to the two *data analysis* stages consisting of exploratory data analysis and confirmatory data analysis (Tukey, 1977). This comprehensive approach to data analysis is described in Sections 1.2.4 and 22.1. Graphical procedures that can be used as *exploratory data analysis* tools for visually discovering the main statistical properties of a data set are presented in Section 22.3 as well as Section 24.2.2 in this chapter. *Confirmatory data analysis* techniques for carrying out hypothesis testing and obtaining rigorous statistical statements about certain statistical properties such as trends are presented throughout the book. These confirmatory tools consist of both parametric models and nonparametric tests. *Parametric models*, such as the intervention models of Chapters 19 and Section 22.4, can be used with data that are not too messy. For example, if there are a few missing observations in a series, the intervention model can be used to estimate these missing values as well as estimate the magnitude of a trend. When the data are very messy, one may have to use *nonparametric methods*. A variety of nonparametric tests for trend detection are given in Section 23.3.

Regression analysis models have a very flexible design and can be used with data that are not evenly spaced over time. In the next section, certain kinds of regression models are described for use as exploratory and confirmatory data analysis tools. A particularly informative regression analysis approach for employment as an exploratory tool for tracing trends is the robust locally weighted regression analysis smooth of Cleveland (1979). This technique is described in detail in Section 24.2.2, while applications of the method are given in that section as well as in 24.3.2.

Table 1.6.4 outlines the three trend analysis methodologies presented in the book. Subsequent to carrying out exploratory data analysis studies and filling in missing observations using the seasonal adjustment method of Section 22.2, intervention analysis is employed in Section 22.4 to describe the impacts of cutting down a forest upon the mean levels of riverflows and water quality. In Section 23.5, water quality applications are used for explaining how exploratory and confirmatory data analysis tools can be employed for studying trends in water quality variables measured in a lake. The purpose of Section 24.3 in this chapter is to present a *general trend analysis methodology for use with water quality time series measured in rivers*. As explained in that section and outlined in Table 24.3.1, the methodology consists of graphical trend studies and trend tests. Different kinds of graphs for observing trends are presented in Sections 22.3 and 24.2.2, while nonparametric trend tests are given in Section 23.3. In Section 24.3, procedures are also described for accounting for the effects of flow or another appropriate covariate upon a given water quality series and eliminating any trend in the flow before its effect upon the water quality series is removed. The *Spearman partial rank correlation test* of Section 23.3.6 provides a powerful test for trend detection in a water quality variable over time when the effects of seasonality or some other factors are partialled out. Water quality data measured in rivers are utilized for illustrating how the trend analysis methodology is applied in practice.

24.2 REGRESSION ANALYSIS

24.2.1 Introduction

Regression analysis constitutes a flexible and highly developed parametric modelling approach which has been applied to virtually every field in which data are measured. A host of books on regression analysis are available including valuable contributions by Mosteller and Tukey (1977), Draper and Smith (1981), Atkinson (1985), and Chambers and Hastie (1992).

Regression models can be designed for modelling many situations, including the type of model in [24.1.1]. In addition to major developments in linear regression models, good progress has been made in nonlinear regression (see, for example, Gallant (1987) and Bates and Watts (1988)).

Consider the case of a linear regression model. When fitting a regression model to a data set, it is recommended to follow the identification, estimation and diagnostic check stages of *model development*, as is also done for all of the time series models presented in this book. A wide variety of well developed procedures are available for constructing linear regression analysis models. Usually, the noise term in a regression model is assumed to be normally independently distributed (NID). If the residuals of a fitted model are not normal, one may be able to rectify the situation by transforming the data using the Box-Cox transformation in [3.4.30]. When the residuals are not independent and hence are correlated, one will have to design a more complicated regression model to overcome this problem or, perhaps, use a completely different type of model such as some kind of stochastic model.

As is also the case for the transfer function noise (TFN) and intervention models of Parts VII and VIII, respectively, a regression model can handle multiple input series. However, recall that in TFN and intervention models, the noise term is correlated and modelled using an ARMA model. In a regression model, the noise term is assumed to be white. Another advantage of a TFN model over a regression model is that the transfer function in a TFN model has an operator in both the numerator and denominator as in [17.2.1] and [17.5.3]. This allows one to handle a wide range of ways in which an input series can effect the output or the response variable, as is discussed at the end of Section 17.2.4 and in Section 19.2.2. Furthermore, this can be done using very few model parameters. In fact, one can think of a TFN model as being a regression model having an autocorrelated noise term rather than white noise. Nonetheless, an advantage of regression analysis over TFN models is that it can be used with data that are not evenly spaced and hence possess missing values.

What is meant by *missing* observations must be explained in more detail, especially for the case of water quality time series (McLeod et al. (1991)). First, consider what missing means with respect to parametric techniques. For many parametric methods, such as the wide variety of time series models given in this book, it is usually assumed that observations are available at equally spaced time intervals. For example, when fitting a periodic autoregressive model in [14.2.1] to average monthly riverflow series, all of the monthly observations across the years must be available. If there is at least one missing observation, the data are no longer evenly spaced due to this missing value. Before fitting the time series model, one must obtain estimates for the missing value or values. Furthermore, for the case of riverflows usually each monthly observation is calculated as a monthly average of average daily flows. Each daily average may be based upon a continuous record taken for that day. Hence, the monthly flow data are often calculated from continuous analogue records.

In contrast to riverflow records, water quality observations usually have quite different meanings in sampling theory. More specifically, most water quality records could be classed as irregular series of *quasi-instantaneous measurements*. This is because each water quality sample takes about 10 to 15 seconds to collect. With such samples, the term *missing data* could refer to a number of situations including:

1. All the unsampled 10 to 15 second periods of the record. This total number is, of course, very large and can be thought of as infinity for practical purposes.
2. Uncollected or lost samples with respect to a specific monitoring objective. For instance, an objective may be to collect one sample per month and if at least one monthly observation is missing the data are irregularly spaced due to this missing value. Another objective may be to have most of the water quality samples taken during times of high flows to produce flow biased data. If no samples are taken for at least one high flow event, then there are missing data.
3. Data missing in the sense of some analytical framework. For example, suppose that the framework chosen is the monthly level. If there is at least one observation per month, then one can say that there are no missing values. However, at a daily level there may be many missing observations.

Another stated characteristic of messy environmental data is that the observations may be significantly affected by *external interventions*. These interventions may be man-induced or natural. An example of a beneficial man-induced intervention is the introduction of tertiary treatment at city sewage plants located in a river basin. This beneficial intervention should cause a step decrease in the phosphorus levels in the river, as shown in Figures 1.1.1 as well as 19.1.1 and explained in Section 19.4.5. On the other hand, uncontrolled industrial development with few environmental controls would cause detrimental impacts upon certain water quality variables in a river. One can cite many other examples of environmental policy and related land use changes which can adversely or beneficially affect water quality. An illustration of a natural intervention is the effect of a forest fire in a river basin upon water quality variables. For example, the resulting lack of forest cover may cause more sediments to be carried and deposited by rivers. However, as a new forest grows back over the years, the flows and water quality variables may slowly revert to their former states, which is the situation for the intervention analysis application presented in Section 19.5.4. In trend analysis one wishes to detect and analyze trends caused by man-made or natural interventions.

The intervention model of Part VIII and Section 22.4 constitutes a flexible type of model that can be used to estimate trends. Regression models can also be designed for estimating the magnitude of trends in a series. One approach is to model all of the components in [24.1.1] employing a regression model. Alternatively, one could describe some of the components in [24.1.1] using regression analysis and the remaining parts utilizing other statistical methods. For instance, a nonparametric trend test (see Section 23.3) could be applied to the residuals of a regression analysis to check for the presence of trends after the covariate and seasonality points have been suitably accounted for using regression analysis (see Section 23.3.5). Brown et al. (1975) employ cumulative sum statistics for tests of change in a regression model structure over time. Within the water resources and environmental engineering literature applications of regression analysis in trend assessment include contributions by Alley (1988), Cunningham and Morton (1983), El-Shaarawi et al. (1983), Loftis et al. (1991), McLeod et al. (1991), Smith and Rose (1991), Reinsel and Tiao (1987), Stoddard (1991), and Whitlatch and Martin (1988).

Esterby and El-Shaarawi (1981a,b) devise a procedure for estimating the *point of change and degree in polynomial regression,* while El-Shaarawi and Esterby (1982) extend the approach for use with a regression model having an autoregressive error process of order one. In environmental engineering problems, the time at which an intervention takes place due to additional

pollution loads and other reasons is often unknown. The approach of Esterby and El-Shaarawi can estimate the time of the intervention (called the point of change) as well as the order of the polynomial regression and associated model parameters both before and after the intervention. Moreover, Esterby (1985) provides a flexible computer program which allows practitioners to implement their versatile regression analysis technique. As an exploratory data analysis tool, the method of Esterby and El-Shaarawi (1981a,b) and El-Shaarawi and Esterby (1982) is useful for detecting the presence and start of the effects of an intervention. Because estimates of the model parameters both before and after the start of the intervention are obtained, their technique can also be considered as a confirmatory data analysis tool. Finally, El-Shaarawi and Delorme (1982) present statistics for detecting a change in a sequence of ordered binomial random variables.

As pointed out earlier in Section 19.2.3, MacNeill (1985) also presents a flexible technique for detecting and modelling the effects of unknown interventions. In particular, he develops a procedure called *adaptive forecasting and estimation using change-detection*. To activate this adaptive procedure, a successively updated change-detection statistic is proposed. The larger the change in the parameters in a regression, exponential smoothing, ARMA or other type of model, the larger is the expected value of the change-detection statistic. Additional research related to MacNeill's change-detection statistic is referred to in Section 19.2.3 under other trend detection techniques. In addition to trend assessment, regression analysis has been extensively utilized for addressing a wide variety of problems arising in water resources and environmental engineering. For example, interesting applications of regression analysis to natural phenomena are provided by Beauchamp et al. (1989), Cleaveland and Durick (1992), Cohn et al. (1992), Duffield et al. (1992), Gunn (1991), Helsel and Hirsch (1992), Keppeler and Ziemer (1990), Kite and Adamowski (1973), Lyman (1992), Millard et al. (1985), Porter and Ward (1991), Potter (1991), See et al. (1992), Simpson et al. (1993), Tasker (1986), Wong (1963), and Wright et al. (1990). Moreover, fuzzy regression analysis (Bardossy, 1990; Kacrzyk and Federizzi, 1992; Tanaka et al., 1982), has been employed in environmental applications (Bardossy et al., 1990, 1992).

Regression models can be used as both exploratory and confirmatory data analysis tools. Section 24.2.2 presents a flexible regression model for visualizing trends in a series at the exploratory data analysis stage. In Section 24.2.3, an example of designing a regression model as a confirmatory tool in an environmental study is presented. Finally, for discussions about the potential pitfalls that one should be aware of when applying regression analysis as well as other kinds of statistical techniques, the reader may wish to refer to the references cited in problem 24.4 at the end of the chapter.

24.2.2 Robust Locally Weighted Regression Smooth

Overview

Suppose that two variables that can be samples are denoted by X and Y. The measurements for these variables are given by (x_i,y_i), $i = 1,2,\ldots,n$. In a scatterplot, the values for the X and Y variables can be plotted as the abscissae and ordinates, respectively. To gain insight into the relationship between X and Y, it is informative to plot some type of smoothed curve through the scatterplot. In the final part of this section, it is explained how smoothed curves can be obtained for a graph of a single time series and also a scatterplot of a time series where values at time $t-k$ are plotted against those at time t. However, for convenience and generality of presentation,

developing a smoothed curve for (x_i,y_i) is now discussed.

A flexible type of smoothing procedure which works well in practice is the *robust locally weighted regression smooth (RLWRS)* developed by Cleveland (1979). Cleveland (1979, 1985), Chambers et al. (1983) and others refer to the general smoothing procedure as LOESS or LOWESS for locally weighted least square regression (when this procedure is iterated robustness is taken into account). Whatever the case, in this chapter the acronym RLWRS is employed. The RLWRS is a member of a set of regression procedures that are commonly referred to as *nonparametric regression* (Stone, 1977). In practice, the RLWRS has been applied to a rich range of problems across many fields. For example, Bodo (1989) and McLeod et al. (1991) have utilized RLWRS for trend assessment of water quality time series, and the RLWRS is also applied to water quality time series in Sections 24.2.2 and 24.3.2 in this chapter. Moreover, Cleveland et al. (1990) have developed a seasonal-trend decomposition procedure based upon the RLWRS.

In essence, the RLWRS is a method for smoothing a scatterplot of (x_i,y_i), $i = 1,2,\ldots,n$, in which the fitted value at x_k is the value of a polynomial fitted to the data using weighted least squares. The weight for (x_i,y_i) is large if x_i is close to x_k and is small if this is not the case. To display graphically the RLWRS on the scatterplot of (x_i,y_i), one plots $(x_i,\hat{y}_i)$ on the same graph as the scatterplot of (x_i,y_i), where $(x_i,\hat{y}_i)$ is called the smoothed point at x_i and $\hat{y}_i$ is called the fitted value at x_i. To form the RLWRS, one simply joins successive smoothed points $(x_i,\hat{y}_i)$ by straight lines. Because a robust fitting procedure is used to obtain the RLWRS, the smoothed points are not distorted by extreme values or other kinds of deviant points.

General Procedure

As explained by Cleveland (1979), the general idea behind his smoothing procedure is as follows. Let W be a weight function which has the following properties:

1. $W(x) > 0$ for $|x| < 1$.
2. $W(-x) = W(x)$.
3. $W(x)$ is a nonincreasing function for $x \geq 0$.
4. $W(x)=0$ for $|x| \geq 1$.

If one lets $0 < f \leq 1$ and r be $(f \cdot n)$ rounded to the nearest integer, the outline of the procedure is as given below. For each x_i, weights, $w_k(x_i)$, are defined for all x_k, $k = 1,2,\ldots,n$, by employing the weight function W. To accomplish this, center W at x_i and scale W so that the point at which W first becomes zero is the rth nearest neighbour of x_i. To obtain the initial fitted value, $\hat{y}_i$, at each x_i a dth degree polynomial is fitted to the data using weighted least squares with weights $w_k(x_i)$. This procedure is called *locally weighted regression.* Based upon the size of the residual $y_i - \hat{y}_i$, a different weight, δ_i, is defined for each (x_i,y_i). In general, large residuals cause small weights while small residuals result in large weights. Because large residuals produce small weights, the effects of extremes tend to be toned down or smoothed, thereby making the procedure *robust.* After replacing $w_k(x_i)$ by $\delta_i w_k(x_i)$, new fitted values are computed using locally weighted regression. The determination of new weights and fitted values is repeated as often as required. All of the foregoing steps taken together are referred to as *robust locally weighted regression.*

In the smoothing procedure, points in the neighbourhood of (x_i,y_i) are used to calculate $\hat{y}_i$. Because the weights $w_k(x_i)$ decrease as the distance of x_k from x_i increases, points whose abscissae are closer to x_i, have a larger effect upon the calculation of $\hat{y}_i$ while further points play a lesser role. By increasing f, the neighbourhood of points affecting $\hat{y}_i$ becomes larger. Therefore, larger values of f tend to cause smoother curves.

In the RLWRS procedure, *local regression* means that regression at a given point is carried out for a subset of nearest neighbours such that the observations closer to the specified point are given larger weights. By taking the size of the residuals into account for obtaining revised weights, robustness is brought into the procedure. Finally, the robust locally weighted regression analysis is carried out for each observation.

Specific Procedure

The procedure presented by Cleveland (1979) for determining the RLWRS is as follows:

1. First the weight function, W, must be specified. Let the distance from x_i to the rth nearest neighbour of x_i be denoted by h_i for each i. Hence, h_i is the rth smallest number among $|x_i - x_j|$, for $j = 1,2,\ldots,n$. For $k = 1,2,\ldots,n$, let

$$w_k(x_i) = W((x_k - x_i)/h_i) \qquad [24.2.1]$$

A possible form for the weight function, is the tricube given by

$$W(x) = (1 - |x|^3)^3 \quad \text{for } |x| < 1$$
$$= 0 \qquad \text{for } |x| > 1 \qquad [24.2.2]$$

2. The second step describes how locally weighted regression is carried out. For each i, determine the estimates, $\hat{\beta}_j(x_i)$, $j = 0,1,\ldots,d$, of the parameters in a polynomial regression of degree d of y_k on x_k. This is fitted using weighted least squares having weight $w_k(x_i)$ for (x_k,y_k). Therefore, the $\hat{\beta}_j(x_i)$ are the values of β_j which minimize

$$\sum_{k=1}^{n} w_k(x_i)(y_k - \beta_0 - \beta_1 x_k - \beta_2 x_k^2 - \cdots - \beta_d x_k^d)^2 \qquad [24.2.3]$$

When using locally weighted regression of degree d, the smoothed point at x_i is $(x_i,\hat{y}_i)$ for which $\hat{y}_i$ is the fitted value of the regression at x_i. Hence,

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(x_i) x_i^j = \sum_{k=1}^{n} r_k(x_i) y_k \qquad [24.2.4]$$

where $r_k(x_i)$ does not depend on y_j, $j = 1,2,\ldots,n$. Cleveland (1979) uses the notation $r_k(x_i)$ to reinforce the fact that the $r_k(x_i)$ are the coefficients for the y_k coming from the regression.

3. Let the bisquare weight function be given by

$$B(x) = (1 - x^2)^2 \quad \text{for } |x| < 1$$
$$= 0, \qquad \text{for } |x| \geq 1 \qquad [24.2.5]$$

Let the residuals for the current fitted values be $e_i = y_i - \hat{y}_i$. The robustness weights are defined by

$$\delta_k = B(e_k/6s) \qquad [24.2.6]$$

where s is the median of the $|e_i|$. As pointed out by Cleveland (1979), other types of weight functions could be used in place of $B(x)$.

4. This step is used to calculate an iteration of robust locally weighted regression. For each i, determine new $\hat{y}_i$ by fitting a dth degree polynomial using weighted least squares having the weight $\delta_k w_k(x_i)$ at (x_k, y_k).

5. Iteratively execute steps 3 and 4 for a total of t' times. The final $\hat{y}_i$ constitute the fitted values for the robust locally weight regression and the $(x_i, \hat{y}_i)$ $i = 1,2, \ldots, n$, form the RLWRS.

Selecting Variables

In order to employ the above procedure, one must specify f, d, t' and W. Cleveland (1979) provides guidelines for doing this. First consider the variable f, where $0 < f \leq 1$, which controls the amount or level of smoothness. As noted earlier, an increase in f causes an increase in the smoothness of the RLWRS. The objective is to select a value of f which is as large as possible to minimize the variability in the smoothed points but without hiding the fundamental pattern or relationship in the data. When it is not certain which value of f to select, setting $f = 0.5$ often produces reasonable results. In practice, one can experiment with two or three values of f and select the one which produces the most informative smooth. Bodo (1989) provides suggestions for selecting f for monthly and lower frequency monitoring data.

Instead of qualitatively selecting one or more values of f, one can estimate f. Based upon the research of Allen (1974), Cleveland (1979) suggests an approach for automatically determining f using a computerized algorithm. The approach begins with the locally weighted regression in step 2. Leaving y_i out of the calculation, for a specified value of f let $\hat{y}_i(f)$ be the fitted value of y_i. A starting value of f_0 for f is chosen by minimizing

$$\sum_{k=1}^{n} (y_k - \hat{y}_k(f))^2 \qquad [24.2.7]$$

Next, using $f = f_0$ the robustness weights in [24.2.6] in step 3 can be determined. Omitting y_i from the calculation and using the robustness weights, let $\hat{y}_i(f)$ be the fitted value at x_i for a given value of f. The next value of f is determined by minimizing

$$\sum_{k=1}^{n} \delta_k (y_k - \hat{y}_k(f))^2 \qquad [24.2.8]$$

Using the latest estimated value of f, the last step can be repeated as many times as are necessary in order to converge to a suitably accurate estimate for f. Depending upon the problem at hand,

this procedure for estimating f may require substantial computational time.

The parameter d is the order of the polynomial that is locally fitted to each point. When $d = 1$, a linear polynomial is specified. This usually results in a good smoothed curve that does not require high computational effort and, therefore, a linear polynomial is commonly used.

The parameter t' stands for the number of iterations of the robust fitting procedure. Based upon experimentation, Cleveland (1979) recommends using $t' = 2$. However, the authors of this book have found that $t' = 1$ is sufficient for most applications.

In the description of the general procedure for RLWRS, four required characteristics of the weight function, $W(x)$, are given. It is also desirable that the weight function smoothly decreases to zero as x goes from 0 to 1. The tricube function in [24.2.2] possesses all of the above stipulated properties. Of course, other appropriate weight functions possessing the above attributes could also be entertained.

Applications

The RLWRS can clearly depict meaningful relationships for the situations described below:

1. *Scatter Plot of X against Y* - The measurements of variables X and Y are given by (x_i,y_i), $i = 1,2,\ldots,n$. In a scatter plot, the values for the X and Y variables are plotted as the abscissae and ordinates, respectively. A RLWRS through the scatter plot can visually display the underlying relationship between X and Y.
2. *Time Series Plot* - In a graph of a time series, one plots the values of the time series x_t at each time t against time $t = 1,2,\ldots,n$. Consequently, the point (x_i,y_i) used in the procedure described above is simply replaced by (t,x_t).
3. *Scatter Plot of a Single Time Series* - In a scatter plot of a variable X, one plots x_{t-k} against x_t in order to see how observations separated by k time lags are related. In the procedure for determining the RLWRS, simply substitute (x_t,x_{t-k}) for (x_i,y_i) in order to obtain the smooth.

Applications are now given to illustrate how RLWRS's can be useful in a time series plot and a scatter plot. The data used in the graphs are water quality data from the Ontario Ministry of the Environment for the Saugeen River at Burgoyne, Ontario, Canada.

As explained in Section 22.3.2, one of the simplest and most informative exploratory data analysis tools is to plot the data against time. Characteristics of the data which are often easily discovered from a perusal of a graph include the detection of extreme values, trends due to known or unknown interventions, dependencies between observations, seasonality, need for a data transformation, nonstationarity and long term cycles.

A time series plot is especially useful for visually detecting the presence or absence of a trend. Figure 24.2.1, for example, shows a graph of logarithmic total nitrates for the Saugeen River against time, where each observation is marked using a cross. The fact that $\lambda = 0$ is written above the graph means that the natural logarithmic transformation from [3.4.30] is invoked. The two horizontal lines plotted in the graph delineate the 95% confidence interval (CI) limits if the series is assumed to be normally independently distributed (NID). The observations that lie far outside the 95% CI in Figure 24.2.1 can be considered as outliers under the assumption that

the data are NID.

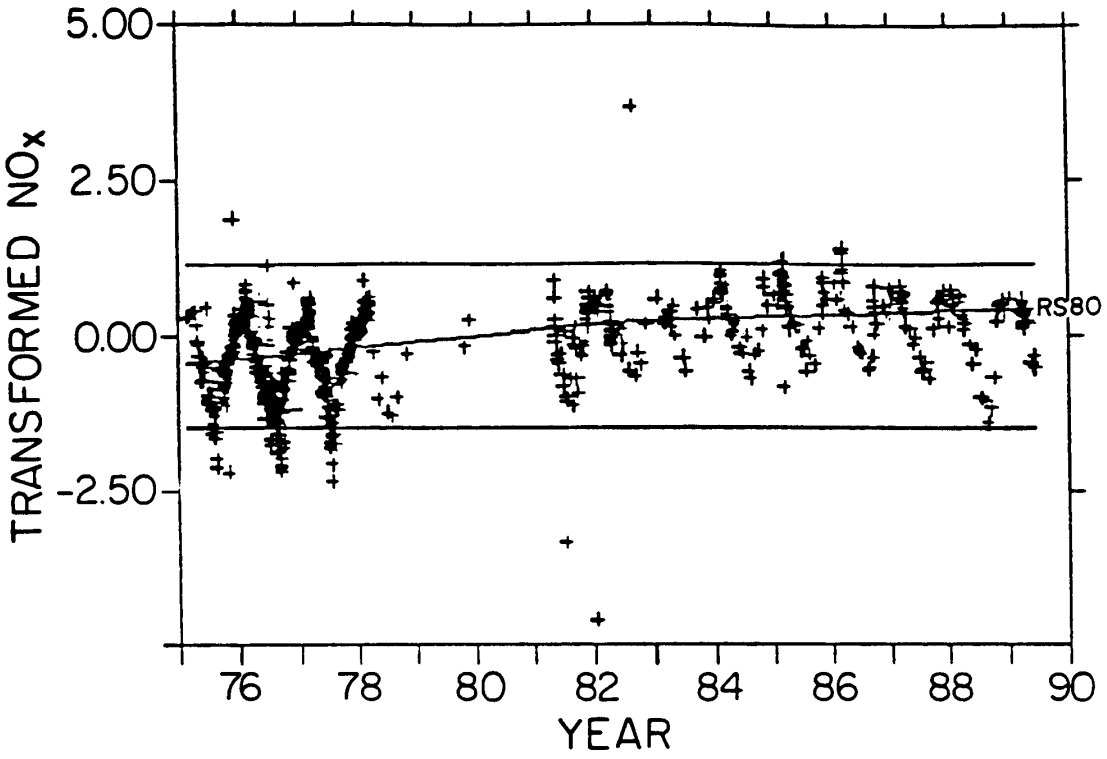

Figure 24.2.1. Graph of logarithmic total nitrates denoted by NO_x (mg/l) against time for the Saugeen River.

In Figure 24.2.1, there is one particularly large observation occurring in 1982. One may wish to examine the records to see if this observation is correct. If the extreme observation were erroneous, one could remove it from the record and thereby not use it in subsequent data analysis. However, the techniques used in the general trend analysis procedure of Section 24.3 are robust or insensitive to outliers. Therefore, this extreme value, and others, are not eliminated from the record.

The line indicating the upward trend through the data is the RLWRS for this time series plot. This robust smooth is referred to in Figure 24.2.1 as RS80 because a value of $f = 0.8$ is employed when plotting the smooth. Hence, the number beside RS is the f value multiplied by 100.

As shown by the dark mass of crosses in the earlier years, more observations were taken at that time. The large gap between many observations, especially in the period from 1979 to 1981, shows that there are time periods during which few measurements were taken and, therefore, the sequence of observations are unequally spaced. The sinusoidal cycle, which is especially pronounced during the first few years, means that the logarithmic total nitrate data are seasonal.

The results of the nonparametric Mann-Kendall trend test (see Section 23.3.2) written below the graph in Figure 24.2.1 confirm that there is an upward trend. This is because the value of the statistic tau calculated using [23.3.5] is positive and the significance level (SL) for this monotonic trend test is close to zero. A small SL (for example, a value less than 0.05) means one should reject the null hypothesis that there is no trend and accept the alternative hypothesis that there is a trend. Alternatively, if the SL level is large and greater than say 0.05, one should accept the null hypothesis that there is no trend. Because the Mann-Kendall test in [23.3.5] or [23.3.1] is designed for employment with nonseasonal data, the result of the test is only a rough indicator for confirming the presence of a monotonic trend in the time series in Figure 24.2.1.

Figure 24.2.2 shows a scatter plot of the logarithmic flows of the Saugeen River (the X variable) against the logarithmic total nitrates (Y variable). The flows are only used for the same times at which the total nitrate measurements are available. Notice that there appears to be a nonlinear function relationship between the flows and the nitrates. To allow the RLWRS to follow this relationship a graph using RS50 is employed. A visual examination of this RLWRS shows that the extreme values do not adversely affect it. The Kendall rank correlation test given at the bottom of Figure 24.2.2 is described in Appendix A23.1. Because tau is positive, there is an upward trend in the scatter plot. The fact that the SL is very small means that the relationship is significant.

24.2.3 Building Regression Models

Overview

Regression analysis constitutes a very general approach to formally modelling statistical data. In fact, regression analysis models can be written in a wide variety of ways and can handle many different situations. When fitting regression models to data sets, one should follow the identification, estimation and diagnostic check stages of model construction, as is done throughout this book for time series models. To illustrate how regression analysis is applied in practice, a case study involving water quality time series is presented.

Lake Erie Water Quality Study

In a particular data analysis study, one should design a specific type of regression model for addressing relevant statistical problems with the data being analyzed. As a brief example of how this is done, consider the statistical data analysis study of water quality time series measured at Long Point Bay in Lake Erie, Ontario, Canada, which is presented in Section 23.5. As summarized in Table 23.5.1, a wide variety of graphical, parametric and nonparametric techniques are utilized for addressing challenging statistical problems. The last item in Table 23.5.1 mentions that regression analysis is employed in the investigation.

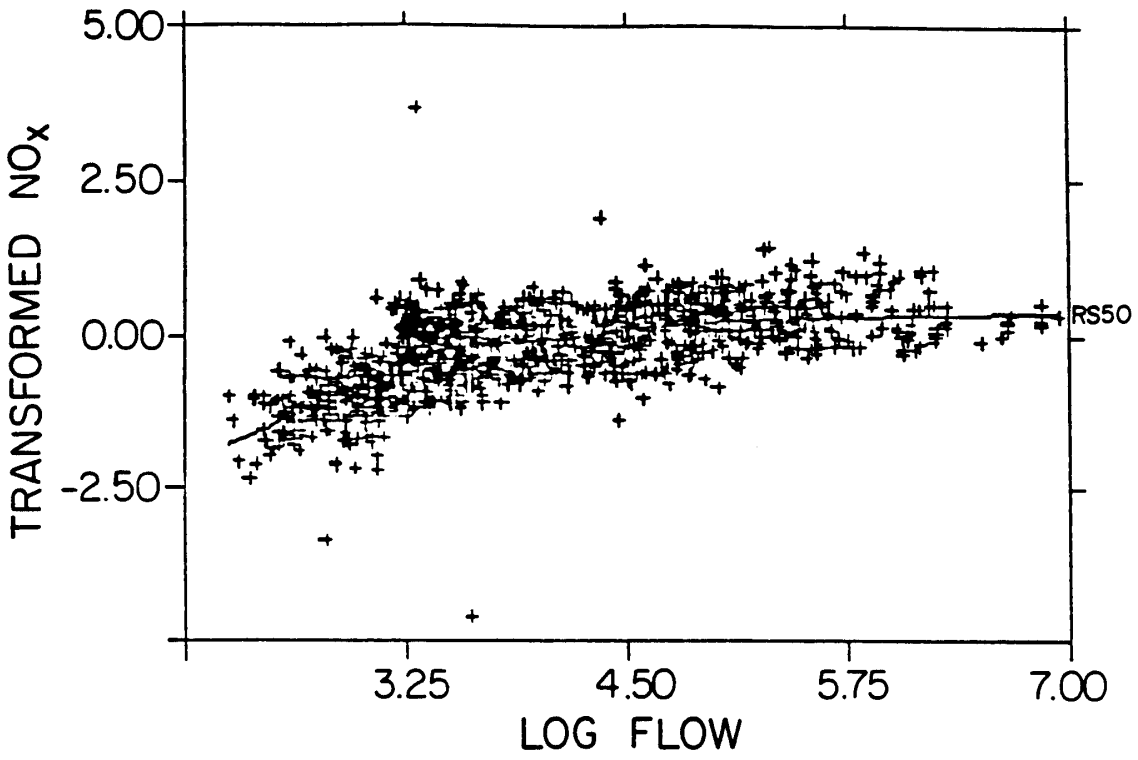

Figure 24.2.2. Scatter plot of logarithmic nitrates against logarithmic flows for the Saugeen River.

Regression Model Design: For the Lake Erie project, a flexible regression model is designed for accomplishing tasks which include determining the best data transformation, ascertaining the components required in a regression model, and estimating both the average monthly and annual values for the series. One can then check for long term trends by examining a smoothed plot of the estimated average annual values.

The most general form of the regression model used in the project is written as

$$y_{ijk}^{(\lambda)} = \mu + \mu_i + \alpha_j + \gamma_{ij} + \beta(x_{ijk} - \bar{x}) + e_{ijk} \qquad [24.2.9]$$

where $i = 1,2,\ldots,n$, denotes the year; $j = 1,2,\ldots,s$, denotes the season when there are s seasons per year (for the monthly Nanticoke data s is usually 9 rather than 12 because often data are not available for the months of January, February and March); $k = 1,2,\ldots,n_{ij}$, denotes the data point in the ith year and jth season for which there is a total of n_{ij} data points; $y_{ijk}^{(\lambda)}$ is the kth data point for a given water quality variable in year i and month j where the λ indicates a Box-Cox transformation (Box and Cox, 1964) which is defined below; μ is the constant term; μ_i is the parameter for the annual effect in the ith year; α_j is the parameter for the seasonal effect in the jth season; γ_{ij} is the interaction term; x_{ijk} is the kth water depth value for the ith year and jth

season; $\bar{x}$ is the mean depth across all of the years and seasons; β is the depth parameter; and e_{ijk} is the error for the kth data point in the ith year and jth season and is normally independently distributed with mean zero and variance σ^2. In order for the model to be identifiable, $\mu_n = 0$, $\alpha_s = 0$, $\gamma_{is} = 0$, $i = 1,2,\ldots,n$, and $\gamma_{nj} = 0$, $j = 1,2,\ldots,s$. Also, if $n_{ij} = 0$ then $\gamma_{ij} = 0$. In addition, $\gamma_{ij} = 0$ if $\sum_{j=1}^{s} n_{ij} = 1$ or $\sum_{i=1}^{n} n_{ij} = 1$. If, for example, $n_{ij} \geq 1$ for all i,j then there are $1 + (n-1) + (s-1) + (n-1)(s-1) = ns$ parameters in the model if one ignores the depth parameter, the transformation parameter λ, and the variance of the error.

Box-Cox Transformation: As is also explained in Section 3.4.5, in order to cause the data to be approximately normally distributed and homoscedastic (i.e., have constant variance), the data can be transformed using a Box-Cox transformation (Box and Cox, 1964) which is defined as

$$y_{ijk}^{(\lambda)} = \begin{cases} \lambda^{-1}(y_{ijk}+c)^{\lambda} - 1 & \lambda \neq 0 \\ \ln(y_{ijk}+c) & \lambda = 0 \end{cases} \qquad [24.2.10]$$

where c is a constant which is usually assigned a magnitude which is just large enough to make all entries in the time series positive. Along with maximum likelihood estimates (MLE's) and standard errors (SE's) for the other model parameters in [24.2.9], one can obtain the MLE of λ and its SE for a given data set. Because it is known that MLE's are asymptotically normally distributed, one can obtain the 95% confidence limits for the MLE of a model parameter such as λ.

Automatic Selection Criteria: A wide variety of statistical procedures are available for selecting the best regression model and making sure that certain modelling assumptions regarding the residuals are satisfied. For example, to choose the most appropriate regression model, one can employ automatic selection criteria such as the AIC and BIC defined in [6.3.1] and [6.3.5], respectively.

R^2 Coefficient: A common criterion for assessing the adequacy of fit of a regression model is the square of the multiple correlation coefficient, denoted by R^2. This statistic reflects the proportion of the total variability which is explained by the fitted regression equation. R^2 has a range between 0 and 1 and the higher the value of R^2, the better is the statistical fit. Consequently, when comparing competing regression models, the one with the highest R^2 value is selected.

Whiteness Tests: To test the adequacy of a fitted model, one can check if one or more assumptions underlying the model residuals are satisfied. In particular, one may wish to ascertain if the residuals are random or uncorrelated. The generalized Durbin-Watson test statistic (Wallis, 1972) provides a test of the null hypothesis that there is no autocorrelation in the residuals of a regression model against the alternative hypothesis that there is significant autocorrelation. More specifically, the test statistic is defined as

$$d_k = \sum_{t=k+1}^{n} (\hat{e}_t - \hat{e}_{t-k})^2 / \sum_{t=1}^{n} \hat{e}_t^2 \qquad [24.2.11]$$

where $\hat{e}_t$ is the residual estimated at time t, n is the length of the residual series, and k is a suitably selected positive integer. Based upon the work of Shively et al. (1990) as well as Ansley et al. (1992), Kohn et al. (1993) develop an algorithm for computing the p-value of the test statistic

in [24.2.11]. Because of this, one can now conveniently execute an hypothesis test for whiteness using the generalized Durbin Watson test statistic. Additionally, the residual autocorrelation function (RACF) defined at lag k as

$$r_k(\hat{e}) = \sum_{t=k+1}^{n} \hat{e}_t \hat{e}_{t-k} / \sum_{t=1}^{n} \hat{e}_t^2 \qquad [24.2.12]$$

is related to the test statistic in [24.2.11] by the relationship

$$r_k(\hat{e}) \approx 1 - \frac{d_k}{2} \qquad [24.2.13]$$

Kohn et al. (1993) furnish a technique for calculating the p-value for the test statistic in [24.2.13] that allows this statistic to also be used as a whiteness test. With nonseasonal data, for example, one may wish to ascertain if r_1 is significantly different from zero. For a seasonal time series, one may also want to examine $r_k(\hat{e})$ for the case where k is the number of seasons per year or some integer multiple of the seasonal length.

The runs test constitutes another procedure for testing whether or not the residuals are white. The runs test is a simple but often effective test of the null hypothesis that a time series is random. Let M denote the median of a time series $z_1, z_2, \ldots, z_n$. If one replaces each z_t by a + or – according as $z_t \leq M$ of $z_t > M$ respectively, then a run is a string of consecutive + or –. The total number of runs, say R, yields a test statistic for randomness. The exact expected number of runs is given by

$$E(R) = 1 + \frac{2n_1 n_2}{n_1 + n_2}, \qquad [24.2.14a]$$

where n_1 is the total number of + and $n_2 = n - n_1$. If there is persistence in the series, the observed number of runs, R, will tend to be less than the expected. On the other hand, for alternating behaviour the number of runs will exceed $E(R)$. The exact variance of R is given by

$$var(R) = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}. \qquad [24.2.14b]$$

Provided that n_1 and n_2 are both greater than 20, the normal approximation can be used to compute the significance level (Swed and Eisenhart, 1943). A two-sided test is used. The Runs Test could be computed about a value other than M but it would have less power. When either $n_1 \leq 20$ or $n_2 \leq 20$ exact formulae given by Swed and Eisenhart (1943) for the probability function of R are used to compute the exact significance level of a two-sided test.

For the case of the regression model in [24.2.9], one would like to ascertain if the residuals, $\hat{e}_{ijk}$, are random. Consequently, in the above test, the z_t series is replaced by the $\hat{e}_{ijk}$ series. The runs test makes no distributional assumption other than independence. When the residuals of a regression model are not random, this would indicate that the fitted model in inadequate.

Analysis of Variance for the Regression: In order to test the statistical significance of the regression model, an ANOVA (analysis of variance) for regression can be executed. For the model in [24.2.9], the null hypothesis is that all parameters (i.e., the μ_i's, α_j's, γ_{ij}'s, and β) are zero except for the mean which would be given by μ when the other parameters are zero. The

alternative hypothesis is that at least one of the parameters, other than the mean, is nonzero.

To calculate the test statistic, various kinds of sums of squares (SS) must be determined. The total SS is

$$SS_{total} = \sum y_{ijk}^2 \qquad [24.2.15]$$

where the $\sum$ refers to summing over all i, j and k. Of course, if the model is fitted to data transformed by a Box-Cox transformation, the SS in [24.2.15] and elsewhere are calculated for the transformed data. The total corrected SS is given by

$$SS_{total\ corrected} = \sum (y_{ijk} - \bar{y})^2 \qquad [24.2.16]$$

where $\bar{y}$ is the overall mean of the y_{ijk}. To calculate the SS of the residuals or the errors, one uses

$$SS_{res} = \sum (y_{ijk} - \hat{y}_{ijk})^2 \qquad [24.2.17]$$

where $\hat{y}_{ijk}$ is the predicted value of y_{ijk} using the fitted regression model in [24.2.9]. The following identity can be used to ascertain the SS for the mean.

$$SS_{mean} = SS_{total} - SS_{total\ corrected} \qquad [24.2.18]$$

Finally, the SS for the regression is given by the identity

$$SS_{reg} = SS_{total} - SS_{mean} - SS_{res} \qquad [24.2.19]$$

To determine the mean square (MS) for a given SS, one divides the SS by the degrees of freedom (DF). The number of degrees of freedom for the regression, denoted by DF_{reg}, is the total number of μ_i, α_j, γ_{ij} and β parameters in [24.2.9] which are not restricted to be zero. If, for instance, $n_{ij} \geq 1$ for all i and j, there are

$$1 + (n-1) + (s-1) + (n-1)(s-1) + 1 = ns + 1$$

degrees of freedom which are due to the mean, μ_i's, α_j's, λ_{ij}'s, and β parameters, respectively. The number of degrees of freedom for the mean correction is simply one while the DF for the SS_{res} is

$$DF_{res} = n - DF_{reg} - 1\ . \qquad [24.2.20]$$

The MS's for the regression and residuals are given by

$$MS_{reg} = \frac{SS_{reg}}{DF_{reg}} \qquad [24.2.21]$$

and

$$MS_{res} = \frac{SS_{res}}{DF_{res}}\ , \qquad [24.2.22]$$

respectively.

The test statistic is given by

$$\hat{F} = \frac{MS_{reg}}{MS_{res}} \qquad [24.2.23]$$

which follows a F distribution with DF_{reg} and DF_{res} degrees of freedom. After calculating the test statistic, one can easily determine the SL in order to ascertain whether or not the null hypothesis should be rejected. For this test, it is assumed that the residuals are normally independently distributed with a mean of zero and a constant variance.

Comparing Other Alternative Models: The ANOVA for the regression compares the simplest possible regression model, for which there is only a mean level, to the full model (FM). In general, one may wish to know whether or not a reduced version of the FM, denoted by RM for reduced model, can describe a data set statistically as well as the more complex FM which contains all of the parameters of the simpler RM. The null hypothesis is that the parameters in the FM which are not contained in the RM are all zero. The alternative hypothesis is that at least one of these parameters is nonzero. If, for example, the null hypothesis were accepted based upon the SL of the test statistic, the RM would adequately model the data and the FM would not be required.

Let $\hat{y}_{ijk}$ and y^*_{ijk} be the values predicted in the regression equations for the FM and RM models, respectively. The SS of the residuals or errors for the FM and RM are given by

$$SS_{res}(FM) = \sum(y_{ijk} - \hat{y}_{ijk})^2 \qquad [24.2.24]$$

and

$$SS_{res}(RM) = \sum(y_{ijk} - y^*_{ijk})^2, \qquad [24.2.25]$$

respectively. The test statistic is then written as

$$\hat{F} = \frac{[SS_{res}(RM) - SS_{res}(FM)]/[DF_{res}(RM) - DF_{res}(FM)]}{SS_{res}(FM)/DF_{res}(FM)} \qquad [24.2.26]$$

where $DF_{res}(FM)$ and $DF_{res}(RM)$ are the numbers of degrees of freedom for the FM and RM, respectively, which are calculated using [24.2.20]. The test statistic in [24.2.26] follows an F distribution with $[DF_{res}(RM)–DF_{res}(FM)]$ and $DF_{res}(FM)$ degrees of freedom. To determine whether or not the null hypothesis should be accepted, the SL for the test statistic can be calculated.

Test for Depth Effect: One may wish to test the hypothesis that an estimated parameter in a fitted regression model is equal to a given constant. If the estimated parameter is $\hat{\beta}_i$, one may wish to test the null hypothesis

$$H_0: \hat{\beta}_i = \beta_i^0$$

where β_i^0 is a constant selected by the investigator. The alternative hypothesis is

$$H_A: \hat{\beta}_i \neq \beta_i^0 .$$

The test statistic is

$$t = \frac{\hat{\beta}_i - \beta_i^0}{SE} \qquad [24.2.27]$$

where t follows a student's t distribution on DF_{res} degrees of freedom calculated using [24.2.19], and SE is the standard error of estimation for $\hat{\beta}_i$. After calculating the SL for the statistic in [24.2.26], one can decide whether or not to accept the null hypothesis.

Estimating Monthly Means: After fitting the regression model in [24.2.9] to a given data set, one can obtain estimates for the average monthly values for those months for which at least some measurements were taken. When the data are transformed using the Box-Cox transformation in [24.2.10], the average monthly values are first calculated for the transformed domain. Letting $\hat{\nu}_{ij}$ stand for the estimate of the average monthly value in year i and month j, then

$$\hat{\nu}_{ij} = \hat{\mu} + \hat{\mu}_i + \hat{\alpha}_j + \hat{\gamma}_{ij} \qquad [24.2.28]$$

where the definitions for the estimated parameters on the right hand side are given in [24.2.9]. The 95% confidence interval for $\hat{\nu}_{ij}$ is $\hat{\nu}_{ij} \pm 1.96SE$. To determine the minimum mean square error (MMSE) estimates of the average monthly means in the untransformed domain, one can use the formulae given by Granger and Newbold (1976) which are discussed in Section 8.2.7. For instance, when $\lambda = 0$ in [24.2.9], the MMSE estimate is

$$\tilde{\nu}_{ij} = \exp[\hat{\nu}_{ij} + \frac{1}{2} var(\hat{\nu}_{ij})] \qquad [24.2.29]$$

To calculate the 95% confidence limits of $\tilde{\nu}_{ij}$, one can replace $\hat{\nu}_{ij}$ by $\hat{\nu}_{ij} + 1.96SE$ and $\hat{\nu}_{ij} - 1.96SE$ in order to determine the upper and lower limits, respectively, for the transformed domain.

Estimating Annual Means: By letting $\hat{\nu}_i$ represent the estimate of the average annual value for year i in the transformed domain, the annual mean for year i can be calculated using

$$\hat{\nu}_i = \frac{1}{s} \sum_{j=1}^{s} \hat{\nu}_{ij} \qquad [24.2.30]$$

where s is the number of seasons for which the ν_{ij} are estimated. The variance of $\hat{\nu}_i$ is determined as

$$var(\hat{\nu}_i) = \frac{1}{s^2} \sum_j \sum_h cov(\nu_{ij} \nu_{ih}) \qquad [24.2.31]$$

The 95% confidence limits for $\hat{\nu}_i$ are $\hat{\nu}_i \pm 1.96SE$ where SE is the square root of the variance in [24.2.28]. To determine the MMSE estimate, $\tilde{\nu}_i$, of the average annual value for year i in the untransformed domain one can employ the formulae of Granger and Newbold (1976). The 95% confidence limits for $\tilde{\nu}_i$ are found by taking the inverse Box-Cox transformation of the 95% confidence limits in the transformed domain. To determine visually if there is a long term trend, one can plot a RLWRS or other kind of smoothed curve through the estimated annual time series. One could also produce a separate graph of the Tukey smooth for the annual values described in Section 22.3.5.

24.3 TREND ANALYSIS METHODOLOGY FOR WATER QUALITY TIME SERIES MEASURED IN RIVERS

24.3.1 Introduction

The collection and analysis of water quality time series are of great import in many regions throughout the world, especially in highly populated and industrialized areas. For example, the Ministry of the Environment within the Canadian province of Ontario operates a spatially and temporally extensive sampling network called the Provincial Water Quality Monitoring Network or simply PWQMN. Approximately one sample per month is collected at over 700 sites, which may be analyzed for up to 60 water quality indicator parameters. In fact, the PWQMN is one of the world's largest water quality sampling networks falling under the umbrella of a single political jurisdiction. Large sums of money are being spent on collecting substantial data through the PWQMN. To make this data meaningful and useful, they must be properly summarized and analyzed. The Ministry of the Environment, as well as many other organizations, are especially interested in detecting and modelling historical trends in PWQMN data. Trend analyses are required for alerting authorities about water quality degradation so that appropriate corrective action can be taken and for evaluating the performance of pollution abatement schemes.

The purpose of this section is to present a general methodology for analyzing trends in water quality time series measured in rivers. When checking for the presence of a trend in a water quality time series, the methodology properly takes into account the effects of riverflows and seasonality upon the water quality observations. Furthermore, the methodology can be used with messy water quality data (see Section 23.1) which may possess undesirable characteristics such as having outliers, non-normality and missing values.

To design the steps presented in the methodology, the authors examined a wide variety of PWQMN water quality time series measured in the Saugeen and Grand Rivers in Southern Ontario, Canada. Based upon the many types of trend analysis problems that arose when analyzing the data, a systematic procedure for studying the time series was developed. Because unforeseen problems were discovered as different kinds of water quality data were analyzed, the trend analysis algorithm was built and improved in an iterative fashion. The final product is a comprehensive and flexible trend analysis methodology for carrying out systematic trend studies of water quality time series.

Within the steps in the methodology, specific graphical, parametric and nonparametric techniques described in Part X of this book are utilized. Although the authors found these techniques to be sufficient for handling all the trend analysis problems they encountered, practitioners and researchers may wish to employ additional specific methods at certain steps in the algorithm. For instance, when looking for basic characteristics of the data by examining graphs of the data, some people may wish to use graphical methods beyond those presented in Section 22.3. Whatever the case, the main steps in the algorithm will remain the same.

In the next section, the steps in the trend analysis methodology are presented and practical applications are employed to demonstrate how the methodology can easily be applied in practice. Although the authors actually applied their methodology to eight PWQMN water quality series plus one waterflow sequence measured at two locations in Southern Ontario, only some representative results are given to explain how the methodology works. Finally, an earlier version of research appearing in this section is provided by McLeod et al. (1991).

24.3.2 Methodology Description

Overview

The methodology given below is valid for use with messy water quality series measured in rivers. In the description, it is assumed that one is ultimately wishing to detect trends at the monthly level. Nonetheless, the methodology can be easily converted for use with other seasonal levels such as quarter-yearly or weekly. To explain the procedure, the NO_x and riverflow data for the Saugeen River at Burgoyne, Ontario, Canada, are employed.

The overall trend analysis study is divided into the two main categories of Graphical Studies and Trend Tests. These two groupings reflect the idea of exploratory and confirmatory data analyses referred to in Sections 1.2.4, 22.1, 23.1 and 24.1. Within the category of Graphical Studies, the following three versions of the water quality series are examined first for trends:

1. *Raw or unadjusted water quality time series*: The given series may be transformed by the Box-Cox transformation in [3.4.30] or [24.2.10] in an attempt to make a non-normal time series become approximately normally distributed (see discussion in Section 3.4.5 and 24.2.3).
2. *Flow-adjusted water quality time series*: This is the time series for which effects of flow upon water quality are removed, as explained in detail below. As mentioned in Section 24.1 just after [24.1.1], in certain situations one may wish to use a covariate series other than flow to adjust the water quality series. If this is the case, replace the word flow by the name of the covariate in the general methodology described in this section.
3. *Detrended-flow-adjusted water quality time series*: After removing trends from the water quantity time series, the influences of flow upon the water quality time series are eliminated in order to produce the detrended-flow-adjusted water data.

Following this, the three average monthly versions of the above three kinds of data are studied using graphical procedures. As noted at the start of this section, one can easily use a seasonal time scale other than monthly. If this is required, replace the word monthly by the designated seasonal category in the description of the methodology.

4. *Mean monthly unadjusted water quality time series.*
5. *Mean monthly flow-adjusted water quality time series.*
6. *Mean monthly detrended-flow-adjusted water quality time series.*

The manner in which these series are calculated is explained below. The main graphical procedure used to examine the six types of water quality data are a trace or time series plot (Section 22.3.2) along with a smoothed curve (called RLWRS in Section 24.2.2) through the plotted data. Finally, under the category of trend tests, the above three types of monthly water quality data are analyzed using trend tests from Chapter 23. Of particular importance is the Spearman partial rank correlation test described in Section 23.3.6 which works extremely well with seasonal data.

The overall trend analysis methodology is summarized in Table 24.3.1. Specific details are presented in Table 24.3.2 for carrying out the Spearman partial rank correlation mentioned opposite d in Table 24.3.1. The steps in the methodology are now explained in detail using the total nitrate (i.e. NO_x) data measured in the Saugeen River at Burgoyne.

Table 24.3.1. Trend analysis methodology for use with water quality time series measured in rivers.

TREND ANALYSIS METHODOLOGY

GRAPHICAL TREND STUDIES

Examine traces along with smoothed curves (i.e. RLWRS's) for the following data sets:

Given Data:

1. Unadjusted water quality time series.
2. Flow-adjusted water quality time series.
3. Detrended-flow-adjusted water quality time series.

Mean Monthly Data:

4. Mean monthly unadjusted water quality time series.
5. Mean monthly flow-adjusted water quality time series.
6. Mean monthly detrended-flow-adjusted water quality time series.

TREND TESTS

For the three mean monthly data sets (i.e. 4, 5 and 6), the following trend tests are carried out:

a. Mann-Kendall (Section 23.3.2).
b. Spearman's rho (Section 23.3.6).
c. Seasonal Mann-Kendall (Section 23.3.2).
d. Spearman partial rank correlation when partialling out seasonality (Section 23.3.6).

To test for seasonality, the Kruskal-Wallis test (Appendix A23.3) and box and whisker graphs (Section 22.3.3) can be used. The tests under c and d are designed for use with seasonal data.

Table 24.3.2. Algorithm for the Spearman partial rank correlation trend test when partialling out seasonality.

ALGORITHM

1) X_t is one of the three monthly series given under 4, 5 and 6 in Table 24.3.1.
2) Test for the presence of seasonality in X_t using
   a. Box and whisker graphs (Section 22.3.3).
   b. Kruskall-Wallis test (Appendix A23.3).
3) If seasonality is not found, use the ordinary Mann-Kendall trend test (Section 23.3.2).
4) When seasonality is present, carry out the Spearman partial rank correlation test of Section 23.3.6 where:
   a. X_t is the series from 1).
   b. $Y_t = t$ where t is the time of the observation.
   c. Z_t is obtained from the ranking of the seasonal effects in the Kruskall-Wallis test (Appendix A23.3).

Graphical Trend Studies

Given Data:

1. Unadjusted water quality time series: As indicated in Table 24.3.1, the first step is to examine a trace along with a RLWRS of the given unadjusted data or the data transformed by a Box-Cox transformation in [3.4.30] or [24.2.10]. A common transformation is to take natural logarithms of the data (i.e. $\lambda = 0$ in [3.4.30]). A graph of the logarithmic total nitrates (NO_x) for the Saugeen River is displayed in Figure 24.2.1. One can easily see the increasing trend over time traced by the RLWRS. Additionally, the results of the Mann-Kendall test at the bottom of the graph also confirm the presence of a trend. However, the SL may not be meaningful because of the high degree of correlation in the data caused by frequent sampling at particular time periods, especially from 1976 to 1978. Moreover, there is also strong seasonality in this data which the Mann-Kendall test cannot properly take into account.

2. Flow-adjusted water quality time series: The question arises as to whether or not a given water quality variable is dependent upon flow. Figure 24.2.2 displays a scatter plot of the logarithmic NO_x series against logarithmic flows. Each flow value is plotted for exactly the same time at which the corresponding NO_x observation is made. As shown by the RLWRS, there is an obvious dependency between NO_x and flow. The value of the Kendall rank correlation test statistic (Appendix A23.1) listed at the bottom of Figure 24.2.2 for the data plotted in this figure, is also significantly large and, therefore, confirms this finding.

Each sample flow value in Figure 24.2.2 is the average daily value for the day on which the corresponding NO_x value was collected. For relatively large rivers such as the Saugeen and

Grand, the discrepancy between instantaneous flow for the exact point in time at which the water quality sample is collected and the mean daily flow, is generally negligible.

Because the water quality values and flows are dependent, one would like to see if a trend is present in the logarithmic water quality time series after the flow effects are removed. To accomplish this, one can examine the residuals of the RLWRS in Figure 24.2.2 . To calculate the residual for each plotted point in Figure 24.2.2, one subtracts the value of the RLWRS at that point. For convenience, a RLWRS using RS50 is used when calculating the residuals. This series is called the flow-adjusted water quality time series. Other approaches for obtaining flow-adjusted water quality time series are discussed in Section 23.3.5.

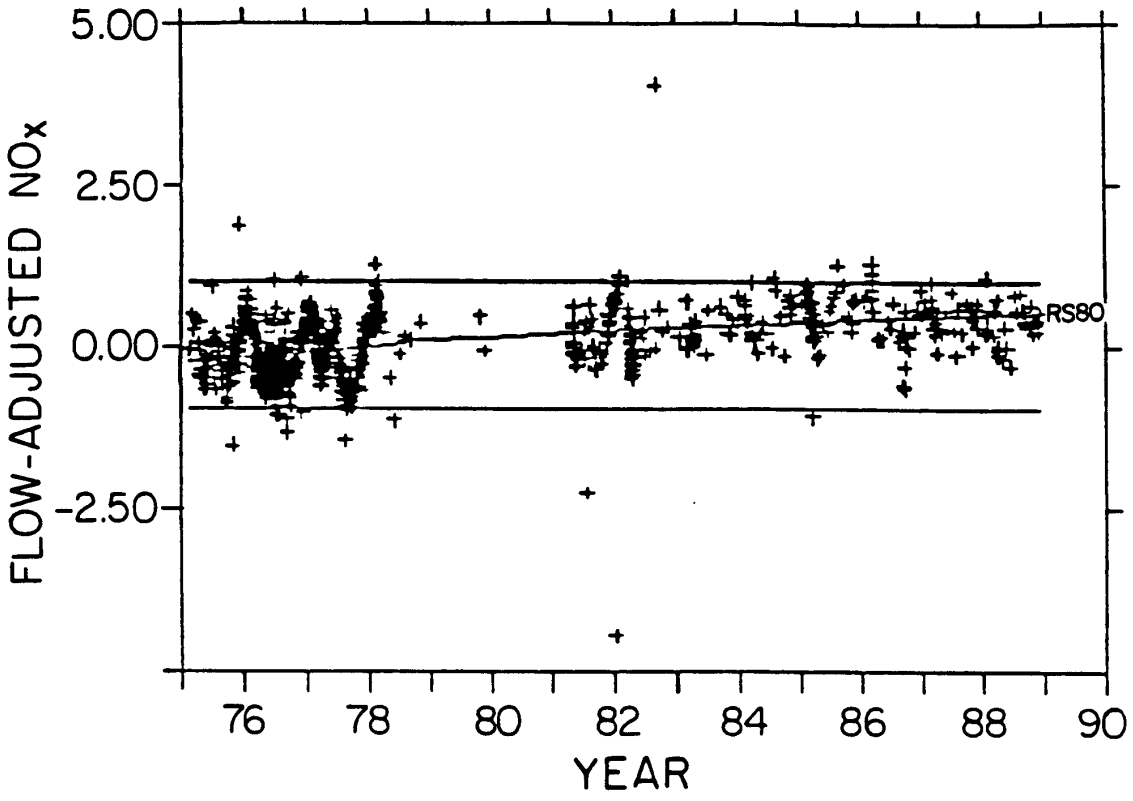

Figure 24.3.1. Graph of the flow-adjusted NO_x series against time for the Saugeen River.

Figure 24.3.1 presents a trace of the flow-adjusted NO_x data for the Saugeen River. Notice that even after flow effects are removed, the RLWRS still shows an obvious upward trend over time.

3. Detrended-flow-adjusted water quality time series: If a trend were present in the flow or covariate series, one would want to remove this trend and then subsequently adjust the water quality time series for detrended flow. In this step, a flexible procedure for obtaining a detrended-flow-adjusted water quality series is described.

Suppose, for now, that the NO_x series really does not have a trend. Flows may cause a trend to appear in the series due to one or both of the following two reasons. First, there may be a real trend in the flows which also causes a trend in the NO_x series. Second, the sampling bias of the flows may cause a trend. Recall that in Figure 24.2.2, each flow is plotted for exactly the same time as the corresponding NO_x observation so that many of the flow observations are not used when producing the flow-adjusted NO_x series. When the complete series of logarithmic Saugeen flows are plotted against time, no trend is present. However, when the logarithmic flows are plotted against time for exactly the same times at which the NO_x values are measured, Figure 24.3.2 shows that the sampling bias has created an obvious trend. To remove the trend from the logarithmic flow series in Figure 24.3.2, one can subtract the RLWRS value from the logarithmic flow series at each time point for which an NO_x observation is available. This residual series is determined for a RLWRS using RS50 to obtain the detrended logarithmic flow series.

Figure 24.3.3 displays a scatter plot of the logarithmic NO_x series against the detrended logarithmic flows. Notice that there is still a dependence between the two series even after the trend due to sampling bias is removed from the logarithmic flows. To obtain the detrended-flow-adjusted NO_x series, one uses the residuals of the RLWRS for the case when RS50 is used to determine the smooth.

For the Saugeen River data, the complete flow record possesses no trend. Because of this, one can state that the trend in the partial flow record plotted in Figure 24.3.2 is due to sampling bias. If the complete record of flows contained a trend, then a trend in the partial flow record (i.e. those flows occurring on the same days at which the water quality samples were collected) would be due to an actual trend in the flows and perhaps also sampling bias.

Figure 24.3.4 shows a graph of the detrended-flow-adjusted data against time. Both the RLWRS and the Mann-Kendall trend test result in this figure show that there is a trend.

Mean Monthly Data

4. Mean monthly unadjusted water quality series: To ascertain the behaviour of the NO_x series at the monthly level, one can examine graphs of the average monthly series. One should keep in mind that the term "average" refers to calculating a mean which may be determined from only a few observations in a given month, each of which is collected in a 10 to 15 second time interval (see discussion on missing values in Section 24.2.1). Figure 24.3.5 shows a graph of the logarithms of the mean monthly NO_x series for the Saugeen River. The RLWRS shows that there is an increasing trend over time. Additionally, there are some months for which no observations are available.

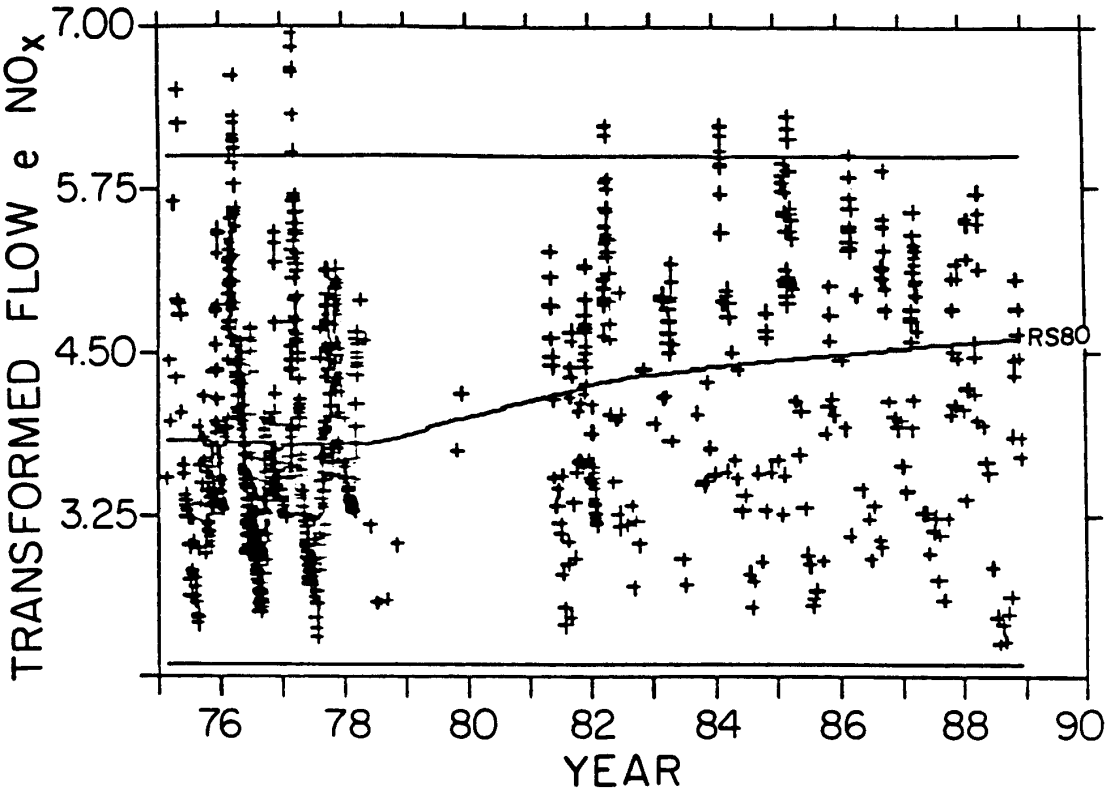

Figure 24.3.2. Logarithmic Saugeen flows against time plotted at exactly the same times at which NO_x observations are available.

5. Mean monthly flow-adjusted water quality time series: Except for the fact that monthly values are used, the logarithmic mean monthly flow-adjusted NO_x series is calculated in exactly the same way as the logarithmic flow-adjusted water quality time series for the given data under item 2. Hence, one determines the residuals of the RLWRS using RS50 fitted to a scatter plot of logarithmic mean monthly NO_x series against the logarithmic mean monthly flows. Figure 24.3.6 displays the graph of the logarithmic mean monthly flow-adjusted water quality series against time, for which there is a striking linear upward trend.

6. Mean monthly detrended-flow-adjusted water quality time series: To eliminate the effects of trends and/or sampling bias in the monthly flows upon the average monthly NO_x series, one can calculate the average monthly detrended-flow-adjusted data. The same procedure followed for determining the flows under item 3 for the given data is also employed here. First, one fits a RLWRS using RS50 to a graph of the logarithmic mean monthly flows against time where the flow observations are only used for months for which NO_x values are available. The

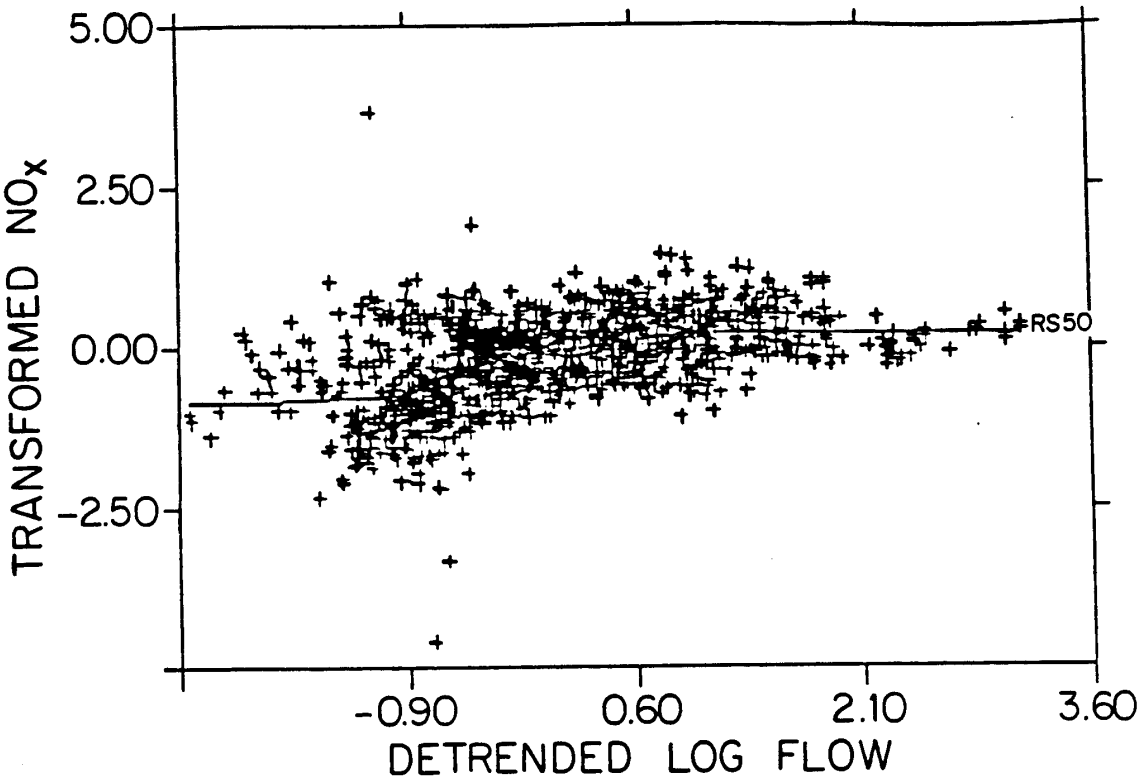

Figure 24.3.3. Scatter plot of the logarithmic NO_x data against detrended logarithmic flows for the Saugeen River.

residuals of the smooth form the logarithmic mean monthly detrended flow series. Second, a RLWRS smooth using RS50 is fitted to a scatter plot of the logarithmic mean monthly NO_x series against the logarithmic mean monthly detrended flow series. The residuals of this smooth constitute the logarithmic average monthly detrended-flow-adjusted NO_x series.

Figure 24.3.7 displays a plot of the logarithmic average monthly detrended-flow-adjusted time series against time. The RLWRS clearly reveals the increasing trend present in this data.

Trend Tests

The four trend tests listed in the bottom half of Table 24.3.1 are applied separately to each of the three types of mean monthly series. For all three series, the four trend tests give exactly the same results. Because of this, representative findings are presented for only one of the three monthly series for explanation purposes in this section.

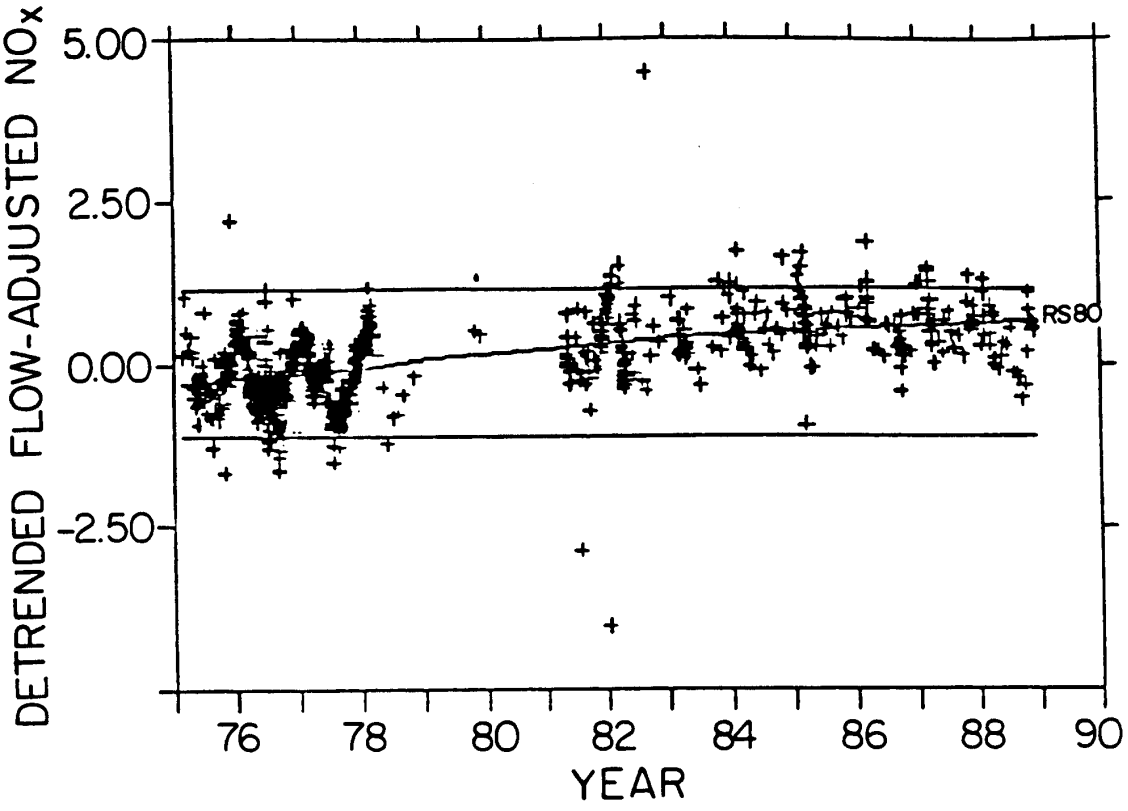

Figure 24.3.4. Graph of detrended-flow-adjusted logarithmic NO_x observations against time for the Saugeen River.

Consider the fifth series which is the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River. Table 24.3.3 presents the results for the four trend tests for this series while Table 24.3.4 gives the average rank value and rank found for each of the twelve seasons used in the Kruskal-Wallis seasonality test. Finally, Figure 24.3.8 displays the box and whisker graph for each of the 12 months for the series being considered. The reader can refer to Section 23.3 for descriptions of the statistical trend tests listed in Table 24.3.3, Appendix A23.3 for a presentation of the Kruskal-Wallis test used in Table 24.3.4, and to Section 22.3.3 for an explanation of box and whisker graphs.

Each of the four trend tests findings in Table 24.3.3 demonstrate that there is a significant trend. In particular, notice that the significance levels are very close to zero for the Mann-Kendall (Section 23.3.2), Spearman's rho (Section 23.3.6), seasonal Mann-Kendall (Section 23.3.2) and the Spearman partial rank correlation (Section 23.3.6) trend tests.

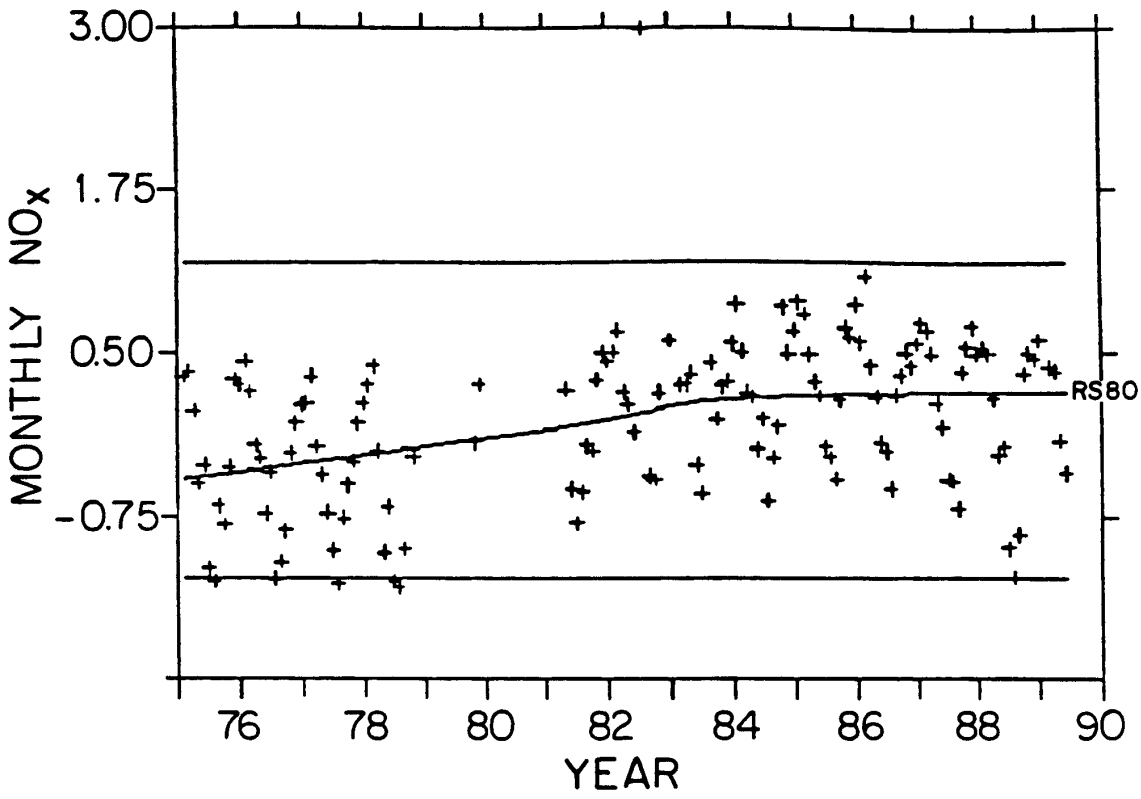

Figure 24.3.5. Graph of the logarithmic average monthly series against time for the Saugeen River.

The test statistic in Table 24.3.4 for the Kruskall-Wallis test has a value of 47.519 and a significance level close to zero. Hence, there is seasonality present in the series. One can also see the cyclic pattern caused by seasonality in the box and whisker graphs for this NO_x series in Figure 24.3.8. Notice, in particular how the median levels change from month to month.

Figure 24.3.8 is an example of what is called a *notched box-and-whisker graph* in Section 22.3.3. The notches on both sides of a box can be used to ascertain if the median in one month is significantly different from another. In particular, when comparing two months, if the median bar in one month overlaps with the notch in the other, and vice versa, then one can argue that the medians for these two months are not significantly different from one another at the 5% significance level. When there are not many data points used to determine a box and whisker plot for a given season, any peculiarities in the plot should be cautiously considered. In Figure 24.3.8, the varying median levels across the months show that the data are seasonal. Notice also for some months that a notch for the mean may extend above an upper hinge or below a lower hinge.

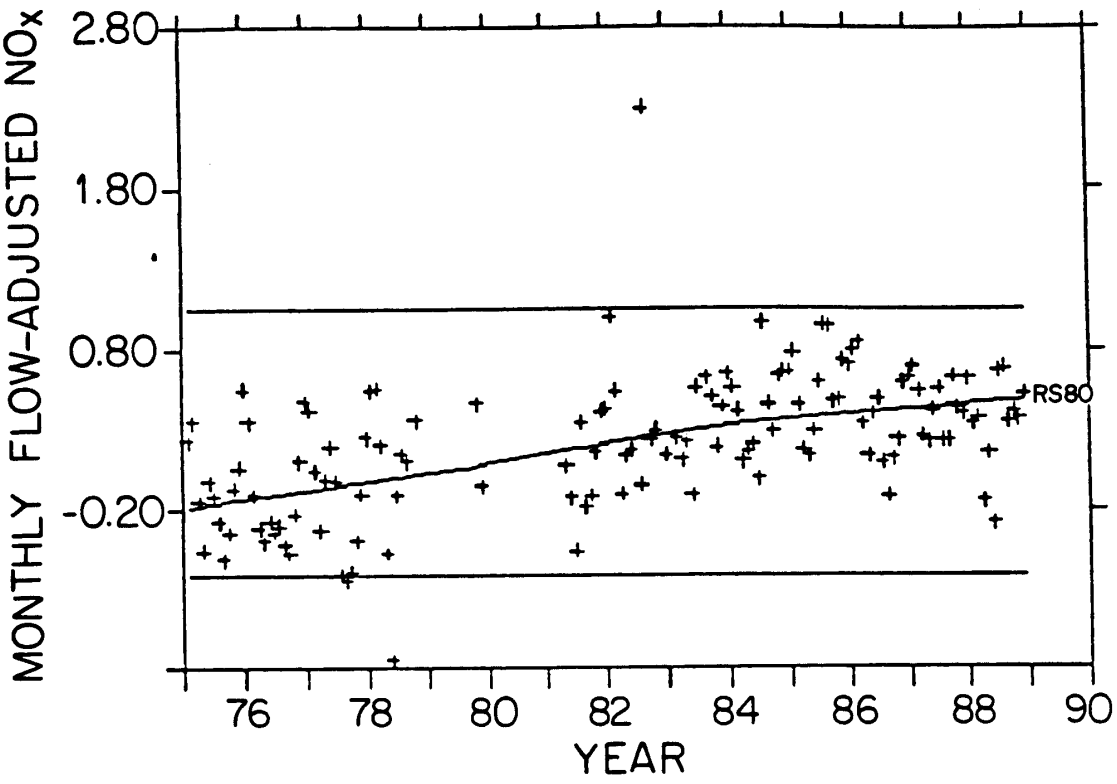

Figure 24.3.6. Graph of the logarithmic mean monthly flow-adjusted NO_x series against time for the Saugeen River.

Because of the importance of the *Spearman partial rank correlation test* for detecting trends in seasonal data, consider the results of this test in more detail. The algorithm for this test is given in Table 24.3.2. In this case, the X_t is the fifth series which is the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River. Under Step 2 of the algorithm in Table 24.3.2, the Kruskall-Wallis test result in Table 24.3.4 as well as the box and whisker graphs of Figure 24.3.8 demonstrate that the data are seasonal. In Table 24.3.4, the seasons are ranked from smallest to largest according to the average rank values for the months. The monthly medians in Figure 24.3.8 can also be compared to obtain the same rankings. Following Step 4 of the algorithm in Table 24.3.2, one lets the Y_t series in the Spearman test be time t while Z_t consists of the seasonal ranks where each month across the years is always given the same rank. By substituting into [23.3.35], one can obtain the Spearman test statistic which has a value of 0.572. Because the SL is almost zero, there is a significant trend over time in the data when the effects of seasonality are partialled out. Since the test statistic is positive, the trend is increasing over time.

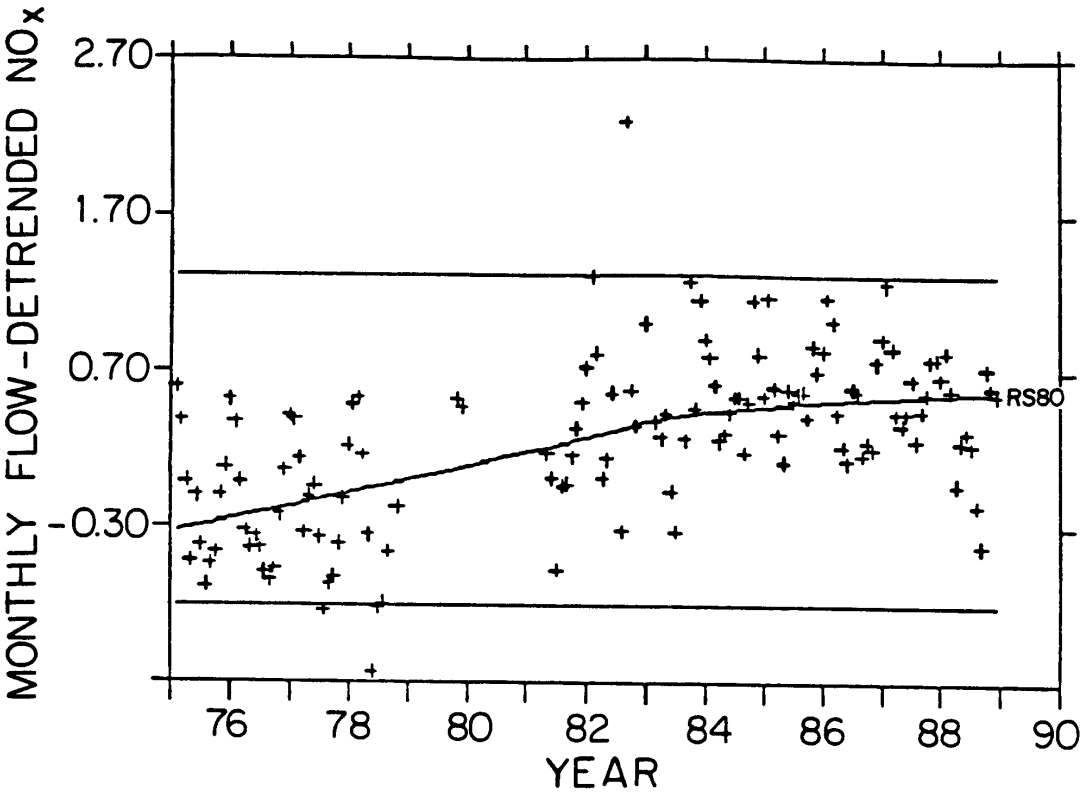

Figure 24.3.7. Graph of the logarithmic average monthly detrended-flow-adjusted data for the Saugeen River.

Table 24.3.3. Trend test results for the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River at Burgoyne.

Trend Tests	Test Statistics	Significance Levels
Mann-Kendall	0.368	0.000
Spearman's Rho	0.542	0.000
Seasonal Mann-Kendall	314	0.000
Spearman Partial Rank Correlation	0.572	0.000

Table 24.3.4. Average ranks and ranks from the Kruskal-Wallis analyses for the 12 months of the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River at Burgoyne.

Months	Sample Sizes	Average Rank Values	Ranks
1	10	89.50	6
2	10	107.70	7
3	11	107.40	7
4	11	92.27	6
5	12	42.00	2
6	12	37.17	1
7	11	46.18	2
8	11	60.82	4
9	12	68.33	5
10	11	54.36	3
11	13	57.92	3
12	11	64.91	4

Test statistic = 47.519 Significance level = 0.000

24.3.3 Summary

A flexible and comprehensive trend analysis methodology is now available for carrying out a systematic study for detecting and modelling trends in water quality series measured in rivers. As summarized in Table 24.3.1, the two main components to the methodology are the Graphical Trend Studies and the Trend Tests. Of particular import and usefulness for analyzing trends in seasonal water quality data is the Spearman partial rank correlation test of Section 23.3.6. When using this test for detecting a trend in a seasonal series for which seasonality is partialled out, the Spearman algorithm of Table 24.3.2 can be utilized. Finally, the overall methodology contains procedures for accounting for the effects of flow upon a given water quality variable.

At various locations in Section 24.3.2, it is noted that one could, if required, use additional graphs and trend tests within the overall trend analysis methodology of Table 24.3.1. For instance, one may wish to employ the adjusted variable Kendall trend test proposed by Alley (1988). However, the authors found that the specific exploratory and confirmatory techniques presented in Part X, readily handled all the situations that arose when examining the water quality series from the Saugeen and Grand Rivers in Southern Ontario.

Another approach to the trend analysis would be to add a deseasonalization step either before or after Step 2 in Table 24.3.1. However, this procedure is not followed here for a number of reasons. First, adjusting the water quality series for flow or some other covariate series may also remove some seasonality. Secondly, an efficient procedure for removing seasonality from a wide variety of messy water quality time series may be very difficult to design. Finally, when seasonality is present in the data sets numbered 4 to 6, a seasonal trend test can be used for checking for the presence of trends in seasonal data (trend tests under c and d in Table 24.3.1).

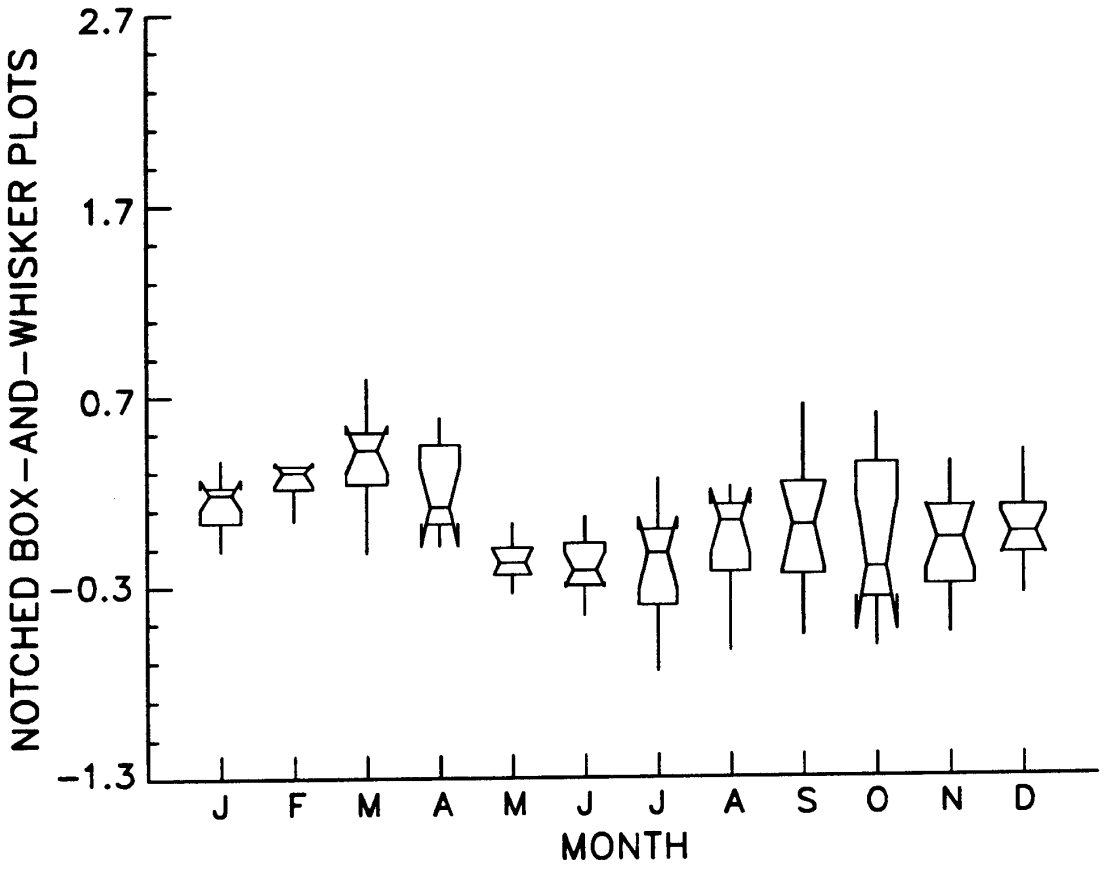

Figure 24.3.8. Box and whisker graphs for the logarithmic mean monthly flow-adjusted NO_x series for the Saugeen River.

As demonstrated by the application to the total nitrates data for the Saugeen River, the methodology of Section 24.3.2 works well in practice. In Table 24.3.5, a summary of the findings for the NO_x data is presented. For both the graphical trend studies and the trend tests, trends are always detected in the versions of the NO_x series that are examined.

Table 24.3.5. Summary of the trend analysis results for the total nitrates data measured in the Saugeen River at Burgoyne.

Data Transformation: Logarithmic

Seasonality: Very strong

Flow-Concentration Relationship: Positive relationship at low flow with relatively constant relationship at higher flow.

Outliers: A few (see Figure 24.2.1). Keep in mind that all techniques used in Table 24.3.1 are robust to outliers.

Trend: All tests indicate a significant trend. Examination of the trace plots suggest that it is largely due to a difference in levels in the data from 1975-1978 and 1982-1989.

Other: Very little data over the period 1979-1981.

As reported elsewhere by McLeod et al. (1991), the authors have used the general trend analysis methodology of Section 24.3.2 with many other water quality time series. In particular, they applied the methodology to the eight PWQMN water quality series (mg/l) listed in Table 24.3.6, as well as riverflow series (m^3/s), for the Saugeen River at Burgoyne and also the Grand River at Dunnville, Ontario. The data at these two sites were selected for study because flow biased monitoring was used. This means that more samples were collected at high flows for the purpose of mass-discharge estimation. Hence, the data collected at these high frequency monitoring sites should contain greater relevant information and their analyses should provide insight into how to analyze both highly monitored and less frequently monitored sites. Representative trend analysis results for the NO_x series for the Saugeen River are employed in Section 24.3.2 for explaining how to apply the methodology.

24.4 CONCLUSIONS

Regression analysis provides a set of flexible statistical tools which can be useful in environmental impact assessment studies as exploratory and confirmatory data analysis methods. A particularly flexible smoothing technique which can be employed as an exploratory data analysis method, for tracing trends in a time series, is the RLWRS of Section 24.2.2. Lewis and Stevens (1991) provide another informative regression approach for drawing a trend curve through a time series. As explained in Section 24.3.2, the RLWRS can also be utilized for removing trends from a series as well as dependent relationships between two series. In Section 24.2.3, a specific case study is used for explaining how a regression analysis model can be

Table 24.3.6. Water quality variables measured in the Saugeen River at Burgoyne and the Grand River at Dunnville, Ontario, Canada.

Water Quality Variables
Ammonia Nitrogen
Total Kjeldahl Nitrogen
Total Nitrates (NO_x)
Filtered Reactive Phosphorus
Total Phosphorus
Suspended Solids
Alkalinity
Conductivity

designed as a confirmatory data analysis technique.

A flexible and comprehensive trend analysis methodology is now available for detecting trends in water quality data measured in rivers. As described in Section 24.3.2, the trend analysis procedure consists of the two main stages of graphical trend studies and trend tests. Specific graphical techniques and statistical trend tests that can be employed in the two main stages are described in detail in Sections 22.3 and Chapter 23, respectively. A particularly powerful trend test for use with seasonal water quality data is the Spearman partial rank correlation test given in Section 23.3.6. Table 24.3.2 presents an algorithm for applying the Spearman partial rank correlation trend test when partialling out seasonality. Application of the trend analysis methodology to water quality series measured in the Saugeen and Grand Rivers demonstrates that the procedure works well in practice.

The overall trend analysis methodology outlined in Table 24.3.1 contains many original developments in environmental impact assessment. Firstly, the RLWRS of Section 24.2.2 is used for calculating flow-adjusted and detrended-flow-adjusted water quality series. Secondly, by employing the detrended-flow-adjusted water quality procedure one can eliminate sampling bias when the entire riverflow series does not possess a trend. As a third contribution, the methodology suggests testing for the presence of seasonality before applying a seasonal trend test such as the seasonal Mann-Kendall test of Section 23.3.2. Simulation studies show that when the data are not seasonal, the seasonal Mann-Kendall test is not as powerful as the Mann-Kendall trend test. Fourthly, the Spearman partial rank correlation test (Section 23.3.6) when partialling out seasonality provides a powerful test for use with seasonal water quality time series. As noted in Section 23.3.6, the Kendall partial rank correlation test cannot be used for this purpose since the distribution of the test statistic is unknown and probably analytically intractable.

PROBLEMS

24.1 Beyond the literature cited in this chapter, locate three other references in which regression analysis is applied to water resources or environmental engineering problems. Briefly explain the purpose of and approach for using regression analysis in each of the papers. Moreover, outline the benefits and drawbacks of employing regression analysis for each of the case studies and suggest how improvements could be made.

24.2 A general approach for trend analysis is given in [24.1.1]. Find a paper in the environmental engineering literature not cited in Chapter 24 in which regression analysis is employed for describing the situation in [24.1.1]. Summarize how the regression analysis is used and explain the advantages and disadvantages of employing the regression procedure as given by the authors of the paper.

24.3 Alley (1988) and also Smith and Rose (1991) describe two basic ways in which regression analysis can be employed in trend assessment. Explain how each of these procedures is carried out and compare their relative strengths and weaknesses.

24.4 Many authors, including Pearson (1897), Huff (1954), Good (1959, 1978), Benson (1965), Wong (1979), Kenny (1982), Wong and DeCoursey (1986), Kite (1989), and Kronmal (1993), discuss statistical fallacies including those arising from the use and abuse of regression analysis. By referring to appropriate literature, clearly explain how spurious correlations and other problems can take place when regression analysis is improperly utilized and how these problems can be overcome.

24.5 Beauchamp et al. (1989) compare regression and time series methods for synthesizing missing streamflow records. After summarizing how they carry out their study, comment upon their findings.

24.6 Outline how the approach of Esterby and El-Shaarawi (1981a,b) and El-Shaarawi and Esterby (1982) works for detecting a point of change in a regression model and estimating the magnitude of the change.

24.7 Concepts from fuzzy set theory have now been incorporated into regression analysis. By referring to the appropriate literature, outline the theory and practice of fuzzy regression analysis and discuss the dividends that can be gained by employing this approach. Describe a hydrological application of fuzzy regression analysis, including a discussion of the insights that are found about the problem being studied.

24.8 As pointed out in Section 24.2.2, Cleveland et al. (1990) have developed a seasonal-trend decomposition procedure based upon the RLWRS (Cleveland, 1979). Outline the main steps in this technique and discuss its advantages and drawbacks. Apply the procedure to a seasonal time series which is of interest to you and be sure to mention any insights which you gain.

24.9 Select two nonseasonal time series between which you feel a meaningful relationship may exist. Plot the RLWRS of Section 24.2.2 on a scatter plot of these two series. Experiment with various values of the smoothing variable, f, where

$0 < f \leq 1$. Discuss the insights that are provided by the graph.

24.10 Carry out the instructions of Problem 24.9 for two seasonal time series.

24.11 Choose a nonseasonal time series that you think may contain a trend. On a time series plot of the series against time, draw the RLWRS. Comment upon the behaviour of any trend that you find in your data set.

24.12 Execute the instructions of Problem 24.11 for the case of a seasonal time series.

24.13 The sample autocorrelation function (ACF), r_k, in [2.5.9] provides a means for quantifying the linear dependence between values in a time series separated by k time lags. To visualize dependence within a single time series, z_t, one can draw a scatter plot of z_t against z_{t-k} along with a RLWRS. Select a nonseasonal time series which is of interest to you and produce a scatter plot and RLWRS of z_t versus z_{t-k} for $k = 1,2,\ldots,7$. Comment upon the type of dependence that you can visually detect in each of the scatter plots. Also, calculate r_k in [2.5.9] for $k = 1,2,\ldots,7$, and compare these results to the visual findings.

24.14 In Section 24.2.3, a specific regression model is designed for modelling water quality time series measured in a lake. Locate a paper in the environmental engineering literature in which the authors employ regression analysis. By using equations when necessary, clearly explain how the authors design, calibrate and check the residual assumptions of their regression model. Describe how the regression analysis assisted the authors in reaching a better understanding about their problem and how their study could be improved.

24.15 For a set of time series that is of direct interest to you and for which it would be appropriate to apply regression analysis, explain how you would design a regression model for studying meaningful relationships among the series. Apply the most appropriate regression model to the data set and check that the residual assumptions are satisfied. Explain the advantages and drawbacks of your approach as well as any surprising results that you uncovered.

24.16 In Section 24.3.2, the Spearman partial rank correlation test defined in Section 23.3.6 is employed for checking for trends after removing seasonality. Using an outline similar to the one given in Table 24.3.2, explain how this test can be utilized for taking into account correlation when testing for the presence of a trend in a time series.

24.17 Table 24.3.1 outlines the trend analysis methodology for use with water quality time series measured in rivers. Beyond the techniques referred to in Section 24.3.2, mention other methods that could be employed with this methodology.

24.18 Select a seasonal water quality time series as well as an accompanying riverflow series that are of interest to you. Carry out the methodology of Section 24.2.2 as well as Section 24.3.2 to check for the presence of trends. Clearly explain all of your steps and comment upon your findings.

24.19 In the trend assessment methodology of Section 24.3.2, it is assumed that monthly data is employed in steps 4 to 6 in Table 24.3.1. Carry out the instructions of Problem 24.18 for the case of quarter-yearly data.

24.20 In Steps 4 to 6 in the trend assessment methodology summarized in Table 24.3.1, it is assumed that monthly data are calculated. Execute the instructions of Problem 24.18 for the situation where weekly data are used in these steps.

REFERENCES

APPLICATIONS OF REGRESSION ANALYSIS

Beauchamp, J. J., Downing, D. J., and Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin*, 25(5):961-975.

Cleaveland, M. K. and Durick, D. N. (1992). Iowa climate reconstructed from tree rings, 1640-1982. *Water Resources Research*, 28(10):2607-2615.

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., and Summers, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research*, 28(9):2353-2363.

Duffield, J. W., Neher, C. J., and Brown, T. C. (1992). Recreational benefits of instream flow: Application to Montana's Big Hole and Bitterroot Rivers. *Water Resources Research*, 28(9):2169-2181.

Gunn, J. (1991). Influences of various forcing variables on global energy balance during the period of intense instrumental observation (1958-1987) and their implications for paleoclimate. *Climatic Change*, 19:393-420.

Keppeler, E. T. and Ziemer, R. R. (1990). Logging effects on streamflow: Water yield and summer low flows at Caspar Creek in Northwestern California. *Water Resources Research*, 26(7):1669-1679.

Kite, G. W. and Adamowski, K. (1973). Stochastic analysis of Lake Superior elevations for computation of relative crustal movement. *Journal of Hydrology*, 18:163-175.

Lyman, R. A. (1992). Peak and off-peak residential water demand. *Water Resources Research*, 28(9):2159-2167.

Millard, S. P., Yearsley, J. R., and Lettenmaier, D. P. (1985). Space-time correlation and its effects on methods for detecting aquatic ecological changes. *Canadian Journal of Fisheries and Aquatic Science*, 42:1391-1400.

Porter, P. S. and Ward, R. C. (1991). Estimating central tendency from uncensored trace level measurements. *Water Resources Bulletin*, 27(4):687-700.

Potter, K. W. (1991). Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment. *Water Resources Research*, 27(5):845-855.

See, R. B., Naftz, D. L., and Qualls, C. L. (1992). GIS-assisted regression analysis to identify sources of selenium in streams. *Water Resources Bulletin*, 28(2):315-330.

Simpson, H. J., Cane, M. A., Herczeg, A. L., Zebiak, S. E. and Simpson, J. H. (1993). Annual river discharge in Southeastern Australia related to El Nino - southern oscillation forecasts of sea surface temperature. *Water Resources Research,* 29(11):3671-3680.

Tasker, G. D. (1986). Accounting for unequal record length and cross correlation in regional regression. In *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium,* July 15-17, 1985, pages 283-290, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

Wong, S. T. (1963). A multivariate statistical model for predicting mean annual flood in New England. *Annals of the Association of American Geographers,* 53(3):298-311.

Wright, K. A., Sendek, K. H., Rice, R. M., and Thomas, R. B. (1990). Logging effects on streamflow: Storm runoff at Caspar Creek in Northwestern California. *Water Resources Research,* 26(7):1657-1667.

BOOKS ON REGRESSION ANALYSIS

Atkinson, A. C. (1985). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications.* Wiley, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Wadsworth and Brooks/Coles, Pacific Grove, California.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis.* Wiley, New York, second edition.

Gallant, A. R. (1987). *Nonlinear Statistical Models.* Wiley, New York.

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

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Massachusetts.

EXPLORATORY DATA ANALYSIS

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Wadsworth, Belmont, California.

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

FUZZY REGRESSION ANALYSIS

Bardossy, A. (1990). Notes on fuzzy regression. *Fuzzy Sets and Systems,* 37(1):65-75.

Bardossy, A., Bogardi, I., and Duckstein, L. (1990). Fuzzy regression in hydrology. *Water Resources Research,* 26(7):1497-1508.

Bardossy, A., Duckstein, L., and Bogardi, I. (1992). Fuzzy nonlinear regression of dose response relationship. *European Journal of Operational Research.*

Kacrzyk, J. and Federizzi, M., editors (1992). *Fuzzy Regression Analysis.* Physica Verlag, Heidelberg.

Tanaka, H., Uejima, S., and Asai, K. (1982). Linear regression analysis with fuzzy model. *IEEE Transactions on Systems, Man and Cybenetics*, SMC-12:903-907.

POINT OF CHANGE IN A REGRESSION MODEL

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37:149-192.

El-Shaarawi, A. H. and Delorme, L. D. (1982). The change-point problem for a sequence of binomial random variables. In El-Shaarawi, A. H. and Esterby, S. R., editors, *Time Series Methods in Hydrosciences*, pages 68-75. Elsevier, Amsterdam, The Netherlands.

El-Shaarawi, A. H. and Esterby, S. R. (1982). Inference about the point of change in a regression model with a stationary error process. In El-Shaarawi, A. H. and Esterby, S. R., editors, *Time Series Methods in Hydrosciences*, pages 55-67. Elsevier, Amsterdam, The Netherlands.

Esterby, S. R. (1985). A program for estimating the point of change and degree in polynomial regression. Technical Report Scientific Series No. 147, Inland Waters Directorate, National Water Research Institute, Burlington, Ontario, Canada.

Esterby, S. R. and El-Shaarawi, A. H. (1981a). Likelihood inference about point of change in a regression regime. *Journal of Hydrology*, 53:17-30.

Esterby, S. R. and El-Shaarawi, A. H. (1981b). Inference about the point of change in a regression model. *Applied Statistics*, 30(3):277-285.

MacNeill, I. B. (1985). Detecting unknown interventions with application to forecasting hydrological data. *Water Resources Bulletin*, 21(5):785-796.

ROBUST LOCALLY WEIGHTED REGRESSION SMOOTH

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125-127.

Bodo, B. A. (1989). Robust graphical methods for diagnosing trend in irregularly spaced water quality time series. *Environmental Monitoring and Assessment*, 12:407-428.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3-33.

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

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Wadsworth, Monterey, California.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 10:1040-1053.

SPURIOUS CORRELATIONS

Benson, M. A. (1965). Spurious correlation in hydraulics and hydrology. *Journal of the Hydraulics Division*, American Society of Civil Engineers (ASCE), 91(HY4):35-42.

Good, I. J. (1959). A classification of fallacious arguments and interpretations. *Technometrics*, 4:125-132.

Good, I. J. (1978). Fallacies, statistical. In Kruskal, W. H. and Tanur, J. M., editors, *International Encyclopedia of Statistics*, Volume 1, pages 337-349. The Free Press, New York.

Huff, D. (1954). *How to Lie with Statistics*. Norton, New York.

Kenney, B. C. (1982). Beware of spurious self-correlations. *Water Resources Research*, 18(4):1041-1048.

Kite, G. (1989). Some statistical observations. *Water Resources Bulletin*, 25(3):483-490.

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisted. *Journal of the Royal Statistical Society*, Series A, 156:379-392.

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society*, London, 60:489-502.

Wong, S. T. (1979). A dimensionally homogeneous and statistically optimal model for predicting mean annual flood. *Journal of Hydrology*, 42:269-279.

Wong, S. T. and DeCoursey, D. G. (1986). More effective development of hydrologic models using dimensional and multivariate analyses. In Shen, H. W., Obeysekera, J. T. B., Yevjevich, V., and DeCoursey, D. G., editors, *Multivariate Analysis of Hydrologic Processes, Proceedings of the Fourth International Hydrology Symposium*, July 15-17, 1985, pages 322-338, Fort Collins, Colorado. Engineering Research Center, Colorado State University.

TECHNIQUES USED WITH REGRESSION ANALYSIS

Alley, W. M. (1988). Using exogeneous variables in testing for monotonic trends in hydrologic time series. *Water Resources Research*, 24(11):1955-1961.

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

Granger, C. W. J. and Newbold, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society*, Series B, 38(2):189-203.

TREND ASSESSMENT USING REGRESSION ANALYSIS

Alley, W. M. (1988). Using exogeneous variables in testing for monotonic trends in hydrologic time series. *Water Resources Research*, 24(11):1955-1961.

Cunningham, R. B. and Morton, R. (1983). A statistical method for the estimation of trend in salinity in the River Murray. *Australian Journal of Soil Research*, 21:123-132.

El-Shaarawi, A. H., Esterby, S. R., and Kuntz, K. W. (1983). A statistical evaluation of trends in the water quality of the Niagara River. *Journal of Great Lakes Research*, 9(2):234-240.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression spline (MARS). *Journal of the American Statistical Association*, 86(416).

Loftis, J. C., Taylor, C. H., Newell, A. D. and Chapman, P. L. (1991). Multivariate trend testing of lake water quality. *Water Resources Bulletin*, 27(3):461-473.

McLeod, A. I., Hipel, K. W., and Bodo, B. A. (1991). Trend analysis methodology for water quality time series. *Environmetrics*, 2(2):169-200.

Reinsel, G. C. and Tiao, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone. *Journal of the American Statistical Association*, 82(397):20-30.

Smith, E. P. and Rose, K. A. (1991). Trend detection in the presence of covariates: Stagewise versus multiple regression. *Environmetrics*, 2(2):153-168.

Stoddard, J. L. (1991). Trends in Catskill stream water quality: Evidence from historical data. *Water Resources Research*, 27(11):2855-2864.

Whitlatch, E. E. and Martin, M. J. (1988). Identification of monthly trends in urban water use. *Water Resources Bulletin*, 24(1):169-174.

WHITENESS TESTS

Ansley, C. F., Kohn, R. and Shirley, T. S. (1992). Computing p-values for the generalized Durbin-Watson and other invariant test statistics. *Journal of Econometrics*, 54:277-300.

Kohn, R., Shively, T. S. and Ansley, C. F. (1993). Computing p-values for the generalized Durbin-Watson statistic and residual autocorrelations in regression. *Applied Statistics*, 42(1):249-269.

Shively, T. S., Ansley, C. F. and Kohn, R. (1990). Fast evaluation of the Durbin-Watson and other invariant test statistics in time series regression. *Journal of the American Statistical Association*, 85:676-685.

Swed, F. S. and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, 14:66-87.

Wallis, K. F. (1972). Testing for fourth-order autocorrelation in quarterly regression equations. *Econometrica*, 40:617-636.