CHAPTER 2

BASIC STATISTICAL CONCEPTS

2.1 INTRODUCTION

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

2.2 TIME SERIES

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

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

riverflow time series. Finally, the *nonparametric trend tests* of Chapter 23, the *regression analysis models* of Chapter 24, and many of the *graphical methods* of Chapter 22 and elsewhere in the book, can be employed with both evenly and unevenly spaced measurements.

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

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

If a polynomial can be fit to a known time series and future entries of the time series can be exactly determined, the time series is said to follow a *deterministic function.* When the future values of a time series cannot be calculated exactly and can be described solely in terms of a probability distribution, the time series is described by a *nondeterministic model* which is usually some kind of statistical or stochastic model. Chronological observations measured from a given

phenomenon form a statistical time series. By knowing the historical values of the widths of the tree rings at a specified site, for example, the range of possible growths for the upcoming years can only be predicted using appropriate probabilistic statements. *This text is involved with modelling natural phenomena which evolve with time according to a probabilistic structure.*

2.3 STOCHASTIC PROCESS

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

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

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

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

population of time series that are defined by the generating stochastic process.

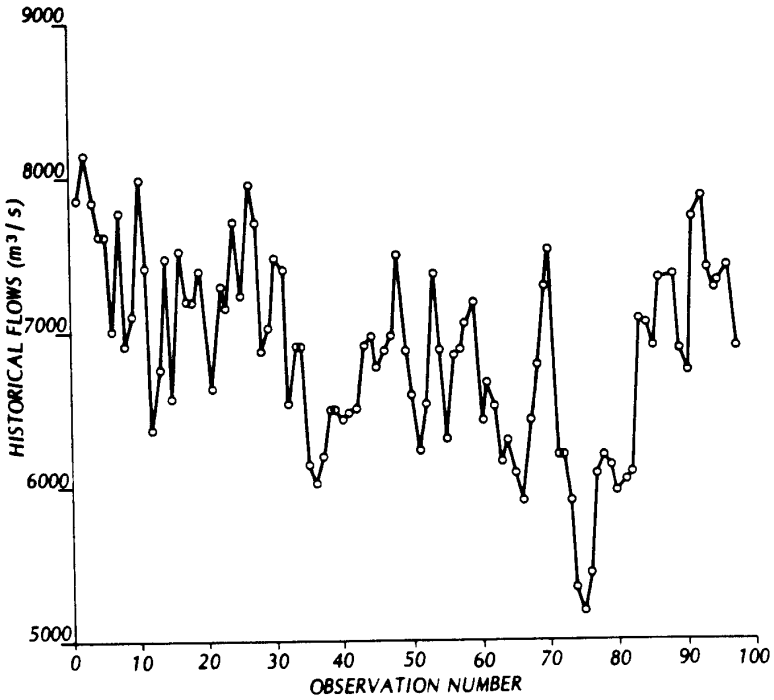

Figure 2.3.1. Annual flows of the St. Lawrence River at Ogdensburg, New York.

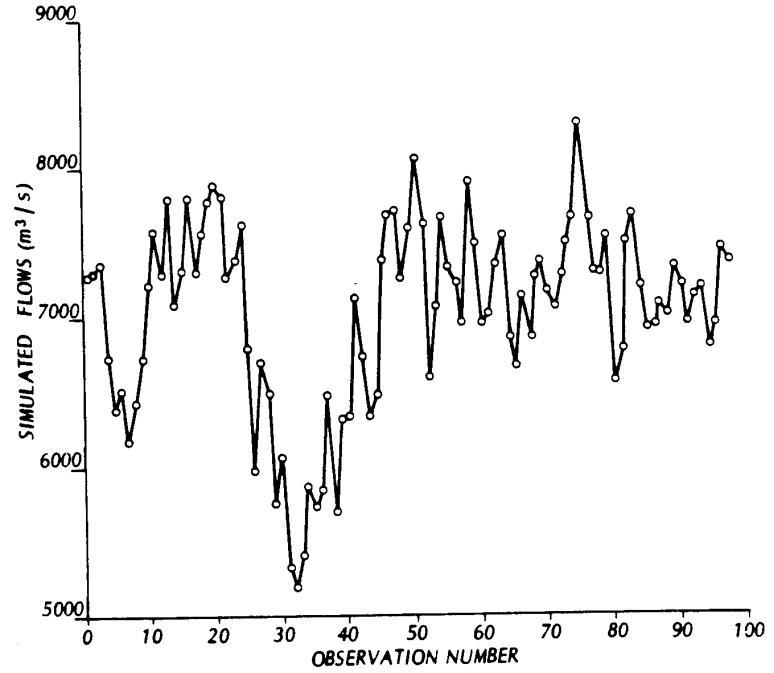

Figure 2.3.2. First simulated sequence of flows for the St. Lawrence River at Ogdensburg, New York.

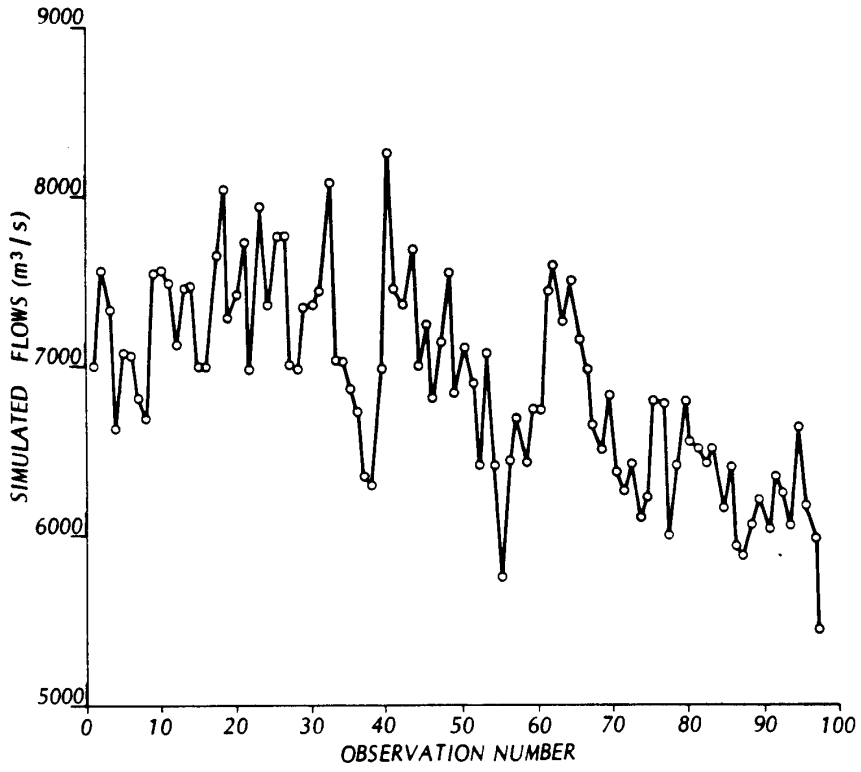

Figure 2.3.3. Second simulated sequence of flows for the St. Lawrence River at Ogdensburg, New York.

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

2.4 STATIONARITY

2.4.1 General Discussion

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

can be fitted to the time series of historical average annual riverflows. Consequently, this infers that the stochastic properties of the complex physical mechanism that produces the observed riverflows, can be represented mathematically by a stationary stochastic process.

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

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

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

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

Within this textbook, *the question of stationary or its antipode, is based upon practical considerations*. When dealing with yearly hydrological and other kinds of natural time series of moderate time spans, it is often reasonable to assume that the process is approximately stationary (Yevjevich, 1972a,b). For example, even though the climate may change slowly over thousands of years, within the time span of a few hundred years the changes in hydrologic time series may be relatively small and therefore these series can be considered to be more or less stationary. If the underlying modelling assumptions are satisfied when a stationary stochastic model is fitted to a nonseasonal series, then these facts validate the assumption of stationarity. When

considering average monthly riverflows, the individual monthly averages may have constant mean values but the means may vary from month to month. Therefore, as explained in Chapters 13 and 14, time series models are employed that reflect the stationarity properties within a given month but recognize the nonstationarity characteristics across all of the months. In other situations, there may be a physical reason for a process to undergo a change in mean level. For example, in 1961 a forest fire in Newfoundland, Canada, devastated the Pipers Hole River basin. In Section 19.5.4, an intervention model is used to model the monthly flows of the Pipers Hole River before and after the fire. The intervention model describes the manner in which the riverflows return to their former patterns as the natural vegetation slowly reverts, over the years, to its condition prior to the fire.

2.4.2 Types of Stationarity

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

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

2.5 STATISTICAL DEFINITIONS

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

2.5.1 Mean and Variance

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

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

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

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

2.5.2 Autocovariance and Autocorrelation

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

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

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

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

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

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

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

Autocovariance and Autocorrelation Matrices

Let the N historical observations be contained in the vector

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

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

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

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

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

The *autocorrelation matrix* is defined by

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

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

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

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

$$var[L_t] = cov[L_t L_t] = E[L_t L_t^T] = E[\mathbf{l}^T(\mathbf{z} - \mu)\{\mathbf{l}^T(\mathbf{z} - \mu)\}^T] = E[\mathbf{l}^T(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T\mathbf{l}]$$

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

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

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

2.5.3 Short and Long Memory Processes

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

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

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

2.5.4 The Sample Autocovariance and Autocorrelation Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

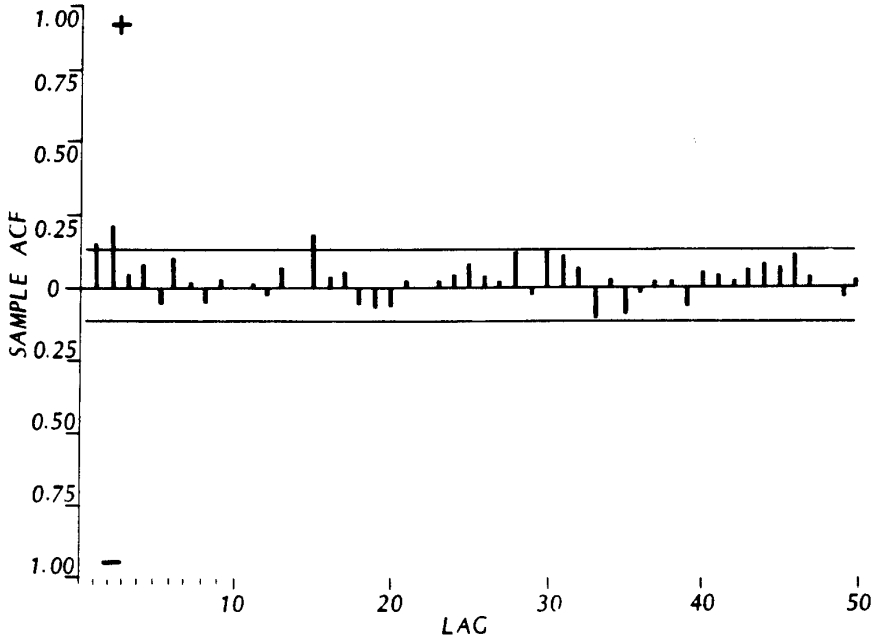

Figure 2.5.1. Sample ACF and 95% confidence limits for the annual temperature data from the English Midlands.

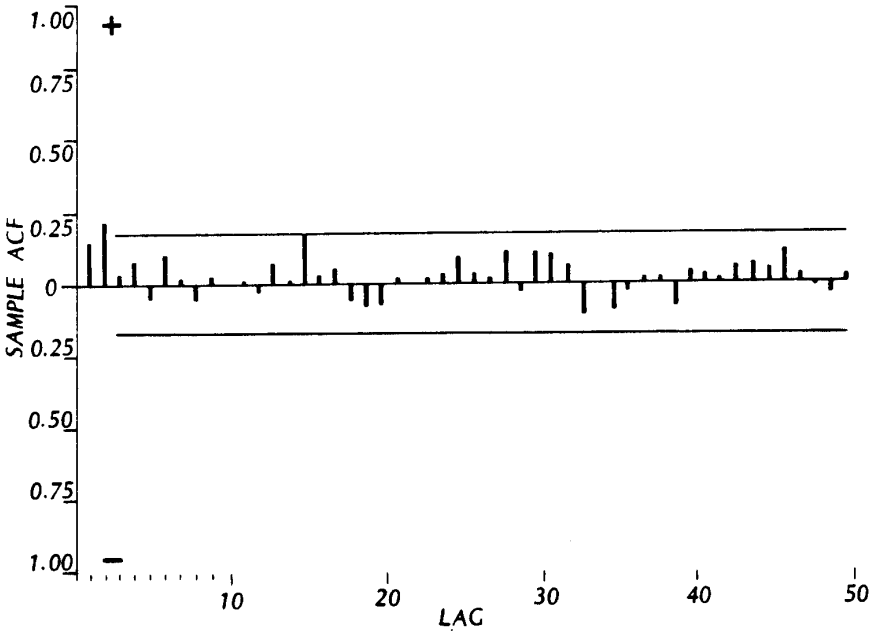

Figure 2.5.2. Sample ACF and 95% confidence limits for the annual temperature data from the English Midlands when ρ_k is zero after lag 2.

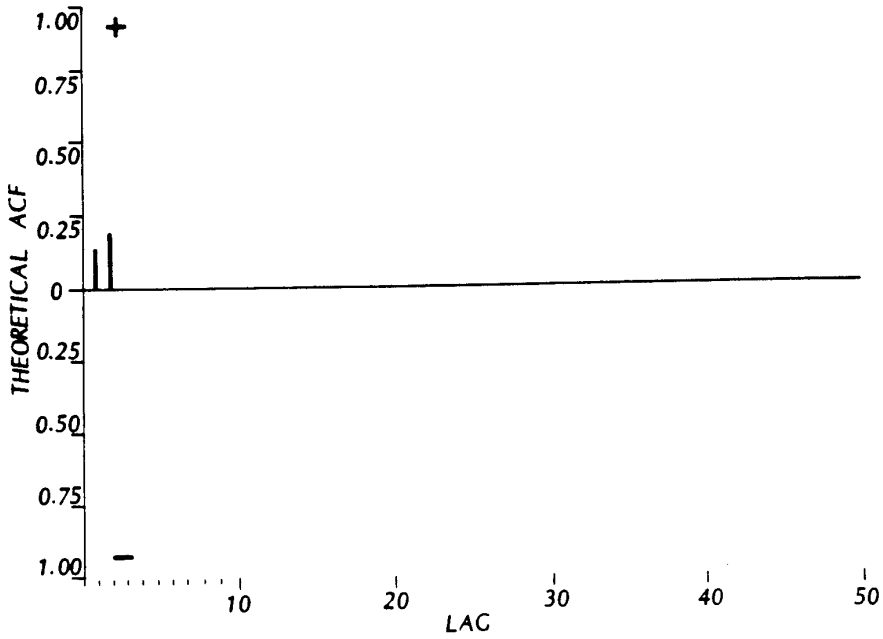

Figure 2.5.3. Theoretical ACF for the model fitted to the temperature data from the English Midlands.

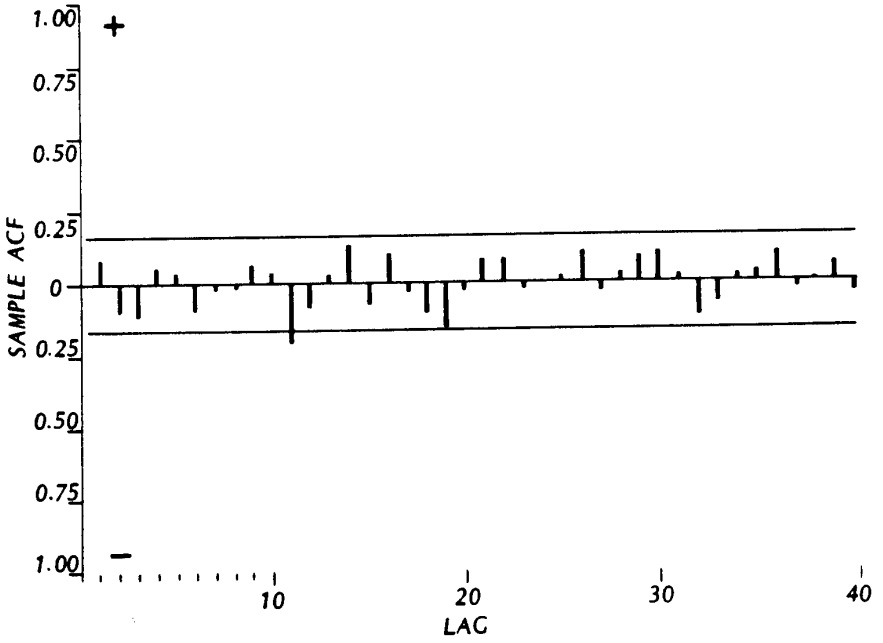

Figure 2.5.4. Sample ACF and 95% confidence limits for the average annual flows of the Rhine River at Basle, Switzerland.

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

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

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

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

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

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

2.5.5 Ergodicity Conditions

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

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

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

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

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

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

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

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

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

2.6 SPECTRAL ANALYSIS

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

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

$$I(f_j) = \frac{2}{N}\left|\sum_{t=1}^{N} z_t \exp(-2\pi i f_j t)\right|$$

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

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

The normalized *cumulative periodogram* is defined by

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

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

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

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

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

where

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

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

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

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

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

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

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

ε	0.01	0.05	0.10	0.25
K_ε	1.63	1.36	1.22	1.02

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

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

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

2.7 LINEAR STOCHASTIC MODELS

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

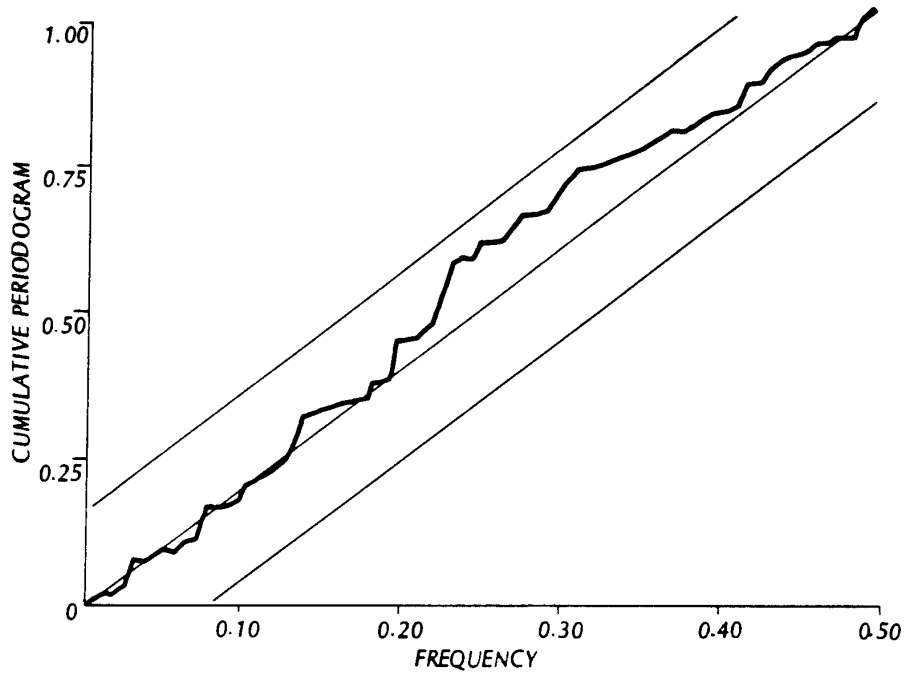

Figure 2.6.1. Cumulative periodogram and 95% confidence limits for the annual Rhine River flows at Basle, Switzerland.

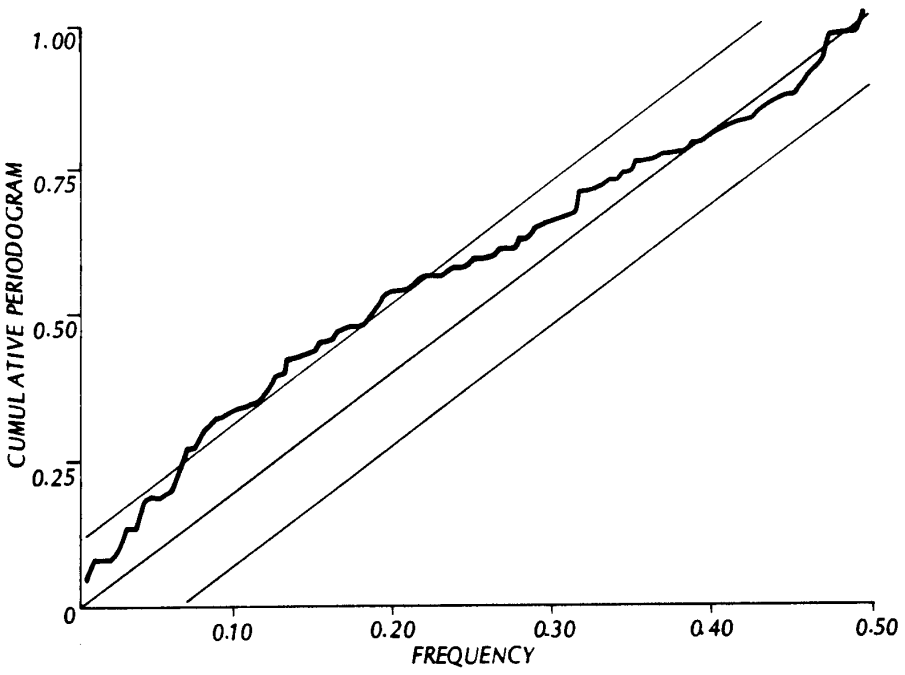

Figure 2.6.2. Cumulative periodogram and 95% confidence limits for the annual temperature data from the English Midlands.

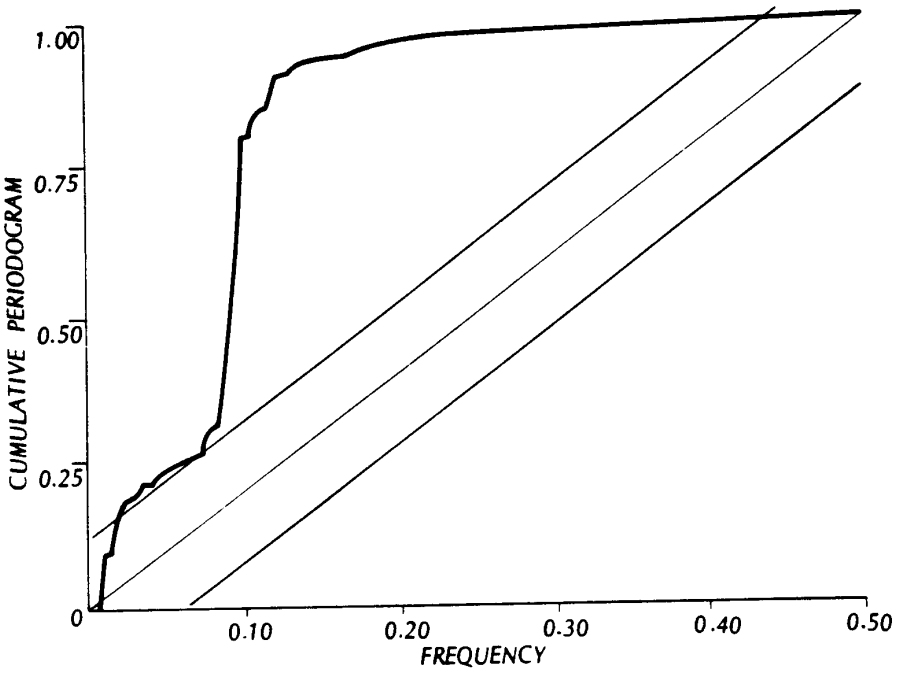

Figure 2.6.3. Cumulative periodogram and 95% confidence limits for the annual sunspot numbers from 1700 to 1960.

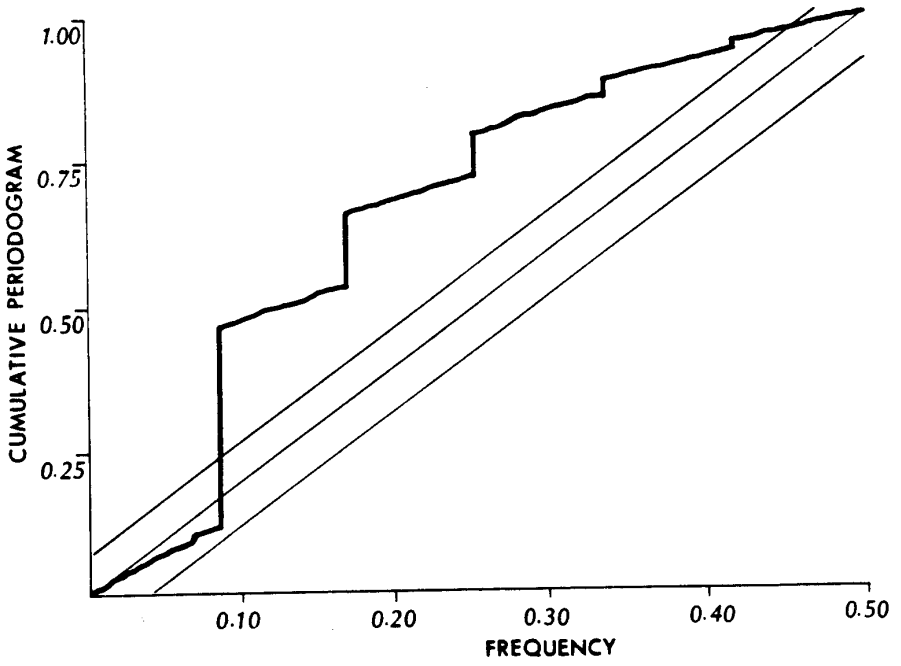

Figure 2.6.4. Cumulative periodogram and 95% confidence limits for the average monthly flows of the Saugeen River.

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

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

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

$$E(a_t) = 0$$

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

and

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

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

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

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

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

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

Consequently, the system is driven by the a_t innovations.

2.8 CONCLUSIONS

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

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

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

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

PROBLEMS

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

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

2.3 Stochastic processes are discussed in Section 2.3. Additionally, in Table 1.4.1 stochastic models are categorized according to the criteria of time (discrete and continuous) and state space (discrete and continuous). By utilizing books referenced in Sections 1.4.3 and 1.6.3,

write down the names of three different kinds of stochastic models for each of the four classifications given in Table 1.4.1.

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

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

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

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

REFERENCES

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

DATA SETS

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

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

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

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

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

HYDROLOGY

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

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

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

STATISTICAL METHODS IN HYDROLOGY

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

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

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

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

STATISTICS

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

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

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

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

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

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

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

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

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

STOCHASTIC PROCESSES

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

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

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

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

TIME SERIES ANALYSIS

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

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

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

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

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

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

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

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

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

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

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

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

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

PART II

LINEAR NONSEASONAL MODELS

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

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

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

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

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

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

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

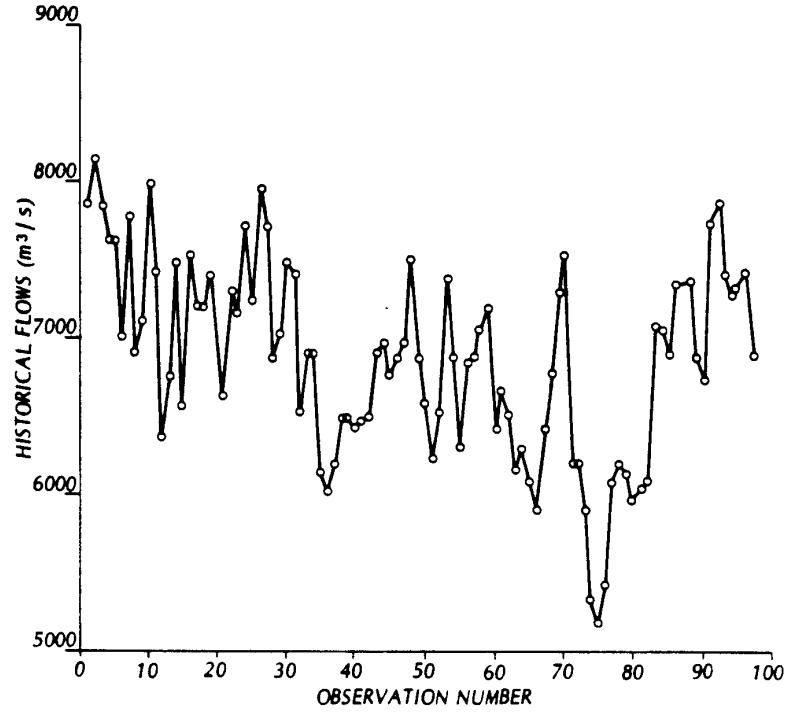

Figure II.1. Annual flows in m^3/s of the St. Lawrence River at Ogdensburg, New York, from 1860 to 1957.

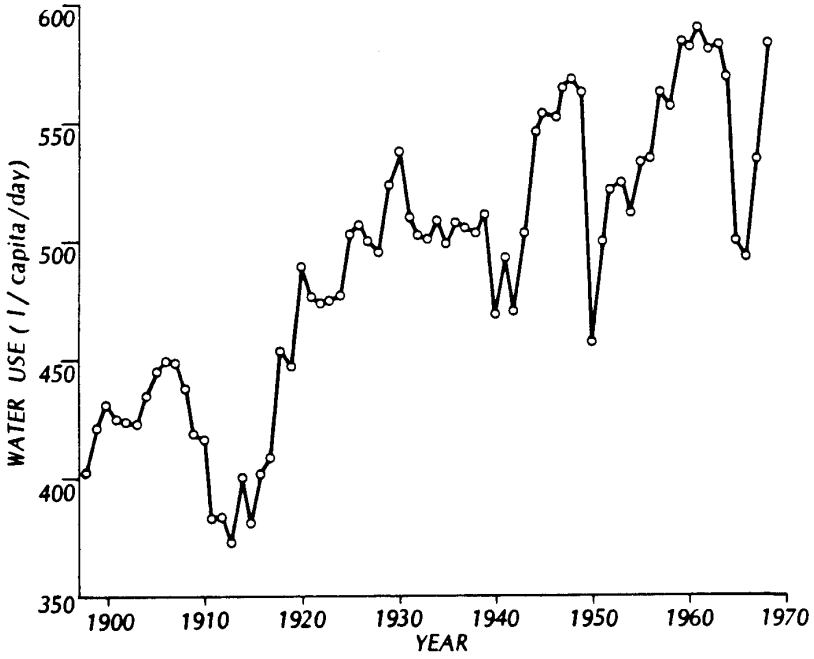

Figure II.2. Annual water use for New York City in litres per capita per day from 1898 to 1970.

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

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

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