1

INTRODUCTION

A *time series* is a sequence of observations taken sequentially in time. Many sets of data appear as time series: a monthly sequence of the quantity of goods shipped from a factory, a weekly series of the number of road accidents, daily rainfall amounts, hourly observations made on the yield of a chemical process, and so on. Examples of time series abound in such fields as economics, business, engineering, the natural sciences (especially geophysics and meteorology), and the social sciences. Examples of data of the kind that we will be concerned with are displayed as time series plots in Figures 2.1 and 4.1. An intrinsic feature of a time series is that, typically, adjacent observations are *dependent*. The nature of this dependence among observations of a time series is of considerable practical interest. *Time series analysis* is concerned with techniques for the analysis of this dependence. This requires the development of stochastic and dynamic models for time series data and the use of such models in important areas of application.

In the subsequent chapters of this book, we present methods for building, identifying, fitting, and checking models for time series and dynamic systems. The methods discussed are appropriate for discrete (sampled-data) systems, where observation of the system occurs at equally spaced intervals of time.

We illustrate the use of these time series and dynamic models in five important areas of application:

1. The *forecasting* of future values of a time series from current and past values.
2. The determination of the *transfer function* of a system subject to inertia—the determination of a dynamic input–output model that can show the effect on the output of a system of any given series of inputs.
3. The use of indicator input variables in transfer function models to represent and assess the effects of unusual *intervention* events on the behavior of a time series.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

4. The examination of interrelationships among several related time series variables of interest and determination of appropriate *multivariate* dynamic models to represent these joint relationships among the variables over time.
5. The design of simple *control schemes* by means of which potential deviations of the system output from a desired target may, so far as possible, be compensated by adjustment of the input series values.

1.1 FIVE IMPORTANT PRACTICAL PROBLEMS

1.1.1 Forecasting Time Series

The use at time t of available observations from a time series to forecast its value at some future time $t + l$ can provide a basis for (1) economic and business planning, (2) production planning, (3) inventory and production control, and (4) control and optimization of industrial processes. As originally described by Holt et al. (1963), Brown (1962), and the Imperial Chemical Industries (ICI) monograph on short term forecasting (Coutie, 1964), forecasts are usually needed over a period known as the *lead time*, which varies with each problem. For example, the lead time in the inventory control problem was defined by Harrison (1965) as a period that begins when an order to replenish stock is placed with the factory and lasts until the order is delivered into stock.

We will assume that observations are available at *discrete*, equispaced intervals of time. For example, in a sales forecasting problem, the sales z_t in the current month t and the sales $z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ in previous months might be used to forecast sales for lead times $l = 1, 2, 3, \ldots, 12$ months ahead. Denote by $\hat{z}_t(l)$ the forecast made at *origin* t of the sales z_{t+l} at some future time $t + l$, that is, at *lead time* l. The function $\hat{z}_t(l)$, which provides the forecasts at origin t for all future lead times, based on the available information from the current and previous values $z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ through time t, will be called the *forecast function* at origin t. Our objective is to obtain a forecast function such that the mean square of the deviations $z_{t+l} - \hat{z}_t(l)$ between the actual and forecasted values is as small as possible *for each lead time* l.

In addition to calculating the best forecasts, it is also necessary to specify their accuracy, so that, for example, the risks associated with decisions based upon the forecasts may be calculated. The accuracy of the forecasts may be expressed by calculating *probability limits* on either side of each forecast. These limits may be calculated for any convenient set of probabilities, for example, 50 and 95%. They are such that the realized value of the time series, when it eventually occurs, will be included within these limits with the stated probability. To illustrate, Figure 1.1 shows the last 20 values of a time series culminating at time t. Also shown are forecasts made from origin t for lead times $l = 1, 2, \ldots, 13$, together with the 50% probability limits.

Methods for obtaining forecasts and estimating probability limits are discussed in detail in Chapter 5. These forecasting methods are developed based on the assumption that the time series z_t follows a *stochastic* model of known form. Consequently, in Chapters 3 and 4 a useful class of such time series models that might be appropriate to represent the behavior of a series z_t, called autoregressive integrated moving average (ARIMA) models, are introduced and many of their properties are studied. Subsequently, in Chapters 6, 7, and 8 the practical matter of how these models may be developed for actual time series data is explored, and the methods are described through the three-stage procedure of tentative

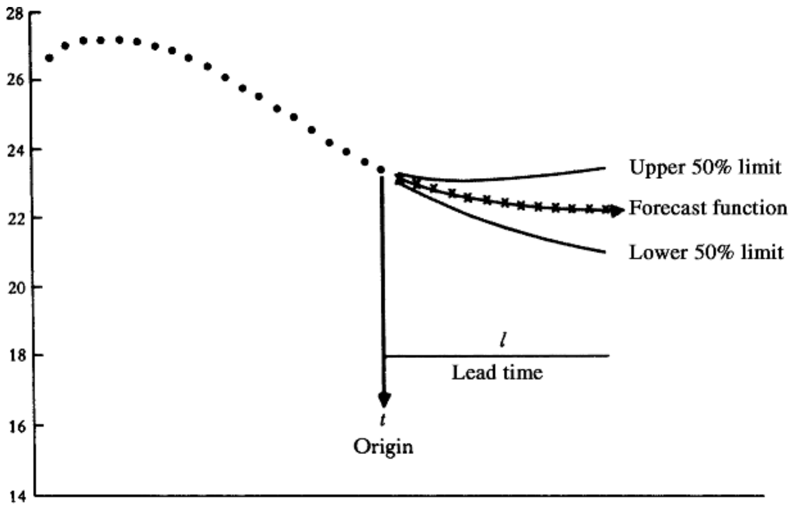

FIGURE 1.1 Values of a time series with forecast function and 50% probability limits.

model identification or specification, estimation of model parameters, and model checking and diagnostics.

1.1.2 Estimation of Transfer Functions

A topic of considerable industrial interest is the study of process dynamics discussed, for example, by Aström and Bohlin (1966, pp. 96–111) and Hutchinson and Shelton (1967). Such a study is made (1) to achieve better control of existing plants and (2) to improve the design of new plants. In particular, several methods have been proposed for estimating the transfer function of plant units from process records consisting of an input time series X_t and an output time series Y_t. Sections of such records are shown in Figure 1.2, where the input X_t is the rate of air supply and the output Y_t is the concentration of carbon dioxide produced in a furnace. The observations were made at 9-second intervals. A hypothetical impulse response function v_j, $j = 0, 1, 2, \ldots$, which determines the *transfer function* for the system through a dynamic linear relationship between input X_t and output Y_t of the form $Y_t = \sum_{j=0}^{\infty} v_j X_{t-j}$, is also shown in the figure as a bar chart. Transfer function models that

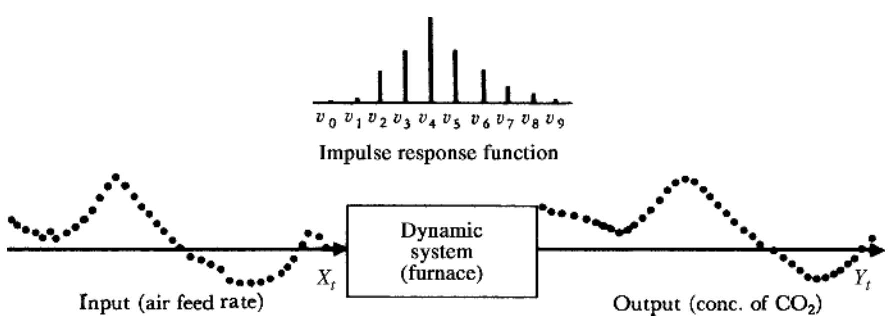

FIGURE 1.2 Input and output time series in relation to a dynamic system.

relate an input process X_t to an output process Y_t are introduced in Chapter 11 and many of their properties are examined.

Methods for estimating transfer function models based on deterministic perturbations of the input, such as step, pulse, and sinusoidal changes, have not always been successful. This is because, for perturbations of a magnitude that are relevant and tolerable, the response of the system may be masked by uncontrollable disturbances referred to collectively as *noise*. Statistical methods for estimating transfer function models that make allowance for noise in the system are described in Chapter 12. The estimation of dynamic response is of considerable interest in economics, engineering, biology, and many other fields.

Another important application of transfer function models is in forecasting. If, for example, the dynamic relationship between two time series Y_t and X_t can be determined, past values of *both* series may be used in forecasting Y_t. In some situations, this approach can lead to a considerable reduction in the errors of the forecasts.

1.1.3 Analysis of Effects of Unusual Intervention Events to a System

In some situations, it may be known that certain exceptional external events, *intervention events*, could have affected the time series z_t under study. Examples of such intervention events include the incorporation of new environmental regulations, economic policy changes, strikes, and special promotion campaigns. Under such circumstances, we may use transfer function models, as discussed in Section 1.1.2, to account for the effects of the intervention event on the series z_t, but where the ''input'' series will be in the form of a simple indicator variable taking only the values 1 and 0 to indicate (qualitatively) the presence or absence of the event.

In these cases, the intervention analysis is undertaken to obtain a quantitative measure of the impact of the intervention event on the time series of interest. For example, Box and Tiao (1975) used intervention models to study and quantify the impact of air pollution controls on smog-producing oxidant levels in the Los Angeles area and of economic controls on the consumer price index in the United States. Alternatively, the intervention analysis may be undertaken to adjust for any unusual values in the series z_t that might have resulted as a consequence of the intervention event. This will ensure that the results of the time series analysis of the series, such as the structure of the fitted model, estimates of model parameters, and forecasts of future values, are not seriously distorted by the influence of these unusual values. Models for intervention analysis and their use, together with consideration of the related topic of detection of outlying or unusual values in a time series, are presented in Chapter 13.

1.1.4 Analysis of Multivariate Time Series

For many problems in business, economics, engineering, and physical and environmental sciences, time series data may be available on several related variables of interest. A more informative and effective analysis is often possible by considering individual series as components of a multivariate or vector time series and analyzing the series jointly. For k-related time series variables of interest in a dynamic system, we may denote the series as $z_{1t}, z_{2t}, \ldots, z_{kt}$, and let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})'$ denote the $k \times 1$ time series vector at time t.

Methods of *multivariate* time series analysis are used to study the dynamic relationships among the several time series that comprise the vector $\boldsymbol{Z}_t$. This involves the development of statistical models and methods of analysis that adequately describe the interrelationships among the series. Two main purposes for analyzing and modeling the vector of time series *jointly* are to gain an understanding of the dynamic relationships over time among the series and to improve accuracy of forecasts for individual series by utilizing the additional information available from the related series in the forecasts for each series. Multivariate time series models and methods for analysis and forecasting of multivariate series based on these models are considered in Chapter 14.

1.1.5 Discrete Control Systems

In the past, to the statistician, the words ''process control'' have usually meant the *quality control techniques* developed originally by Shewhart (1931) in the United States (see also Dudding and Jennet, 1942). Later on, the sequential aspects of quality control were emphasized, leading to the introduction of *cumulative sum charts* by Page (1957, 1961) and Barnard (1959) and the *geometric moving average* charts of Roberts (1959). Such basic charts are frequently employed in industries concerned with the manufacture of discrete ''parts'' as one aspect of what is called *statistical process control* (SPC). In particular (see Deming, 1986), they are used for continuous monitoring of a process. That is, they are used to supply a continuous screening mechanism for detecting assignable (or special) causes of variation. Appropriate display of plant data ensures that significant changes are quickly brought to the attention of those responsible for running the process. Knowing the answer to the question ''*when* did a change of this particular kind occur?'' we may be able to answer the question ''*why* did it occur?'' Hence a continuous incentive for process stabilization and improvement can be achieved.

By contrast, in the process and chemical industries, various forms of *feedback and feedforward* adjustment have been used in what we will call *engineering process control* (EPC). Because the adjustments made by engineering process control are usually computed and applied automatically, this type of control is sometimes called *automatic process control* (APC). However, the *manner* in which these adjustments are made is a matter of convenience. This type of control is necessary when there are inherent *disturbances* or *noise* in the system inputs that are impossible or impractical to remove. When we can measure fluctuations in an input variable that can be observed but not changed, it may be possible to make appropriate compensatory changes in some other control variable. This is referred to as *feedforward control*. Alternatively, or in addition, we may be able to use the deviation from target or ''error signal'' of the output characteristic itself to calculate appropriate compensatory changes in the control variable. This is called *feedback control*. Unlike feedforward control, this mode of correction can be employed even when the source of the disturbances is not accurately known or the magnitude of the disturbance is not measured.

In Chapter 15, we draw on the earlier discussions in this book, on time series and transfer function models, to provide insight into the statistical aspects of these control methods and to appreciate better their relationships and different objectives. In particular, we show how some of the ideas of feedback control can be used to design simple charts for *manually adjusting* processes. For example, the upper chart of Figure 1.3 shows hourly measurements of the viscosity of a polymer made over a period of 42 hours. The viscosity is to be controlled about a target value of 90 units. As each viscosity measurement comes to hand, the process operator uses the nomogram shown in the middle of the figure to compute the adjustment to be made in the manipulated variable (gas rate). The lower chart of Figure 1.3 shows the adjustments made in accordance with the nomogram.

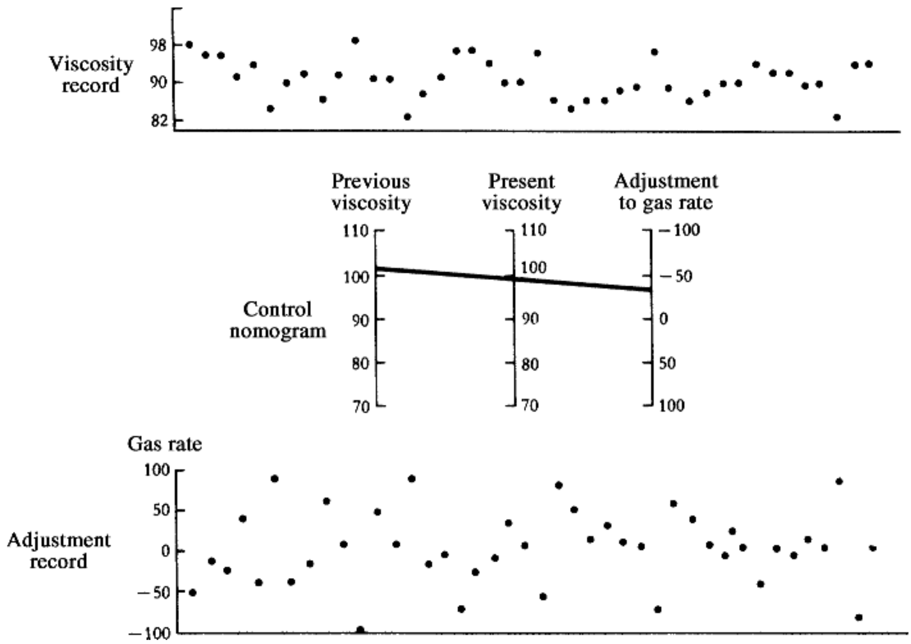

FIGURE 1.3 Control of viscosity. Record of observed viscosity and of adjustments in gas rate made using nomogram.

1.2 STOCHASTIC AND DETERMINISTIC DYNAMIC MATHEMATICAL MODELS

The idea of using a mathematical model to describe the behavior of a physical phenomenon is well established. In particular, it is sometimes possible to derive a model based on physical laws, which enables us to calculate the value of some time-dependent quantity nearly exactly at any instant of time. Thus, we might calculate the trajectory of a missile launched in a known direction with known velocity. If exact calculation were possible, such a model would be entirely *deterministic*.

Probably no phenomenon is totally deterministic, however, because unknown factors can occur such as a variable wind velocity that can throw a missile slightly off course. In many problems, we have to consider a time-dependent phenomenon, such as monthly sales of newsprint, in which there are many unknown factors and for which it is not possible to write a deterministic model that allows exact calculation of the future behavior of the phenomenon. Nevertheless, it may be possible to derive a model that can be used to calculate the *probability* of a future value lying between two specified limits. Such a model is called a probability model or a *stochastic model*. The models for time series that are needed, for example, to achieve optimal forecasting and control, are in fact stochastic models. It is necessary in what follows to distinguish between the probability model or stochastic process, as it is sometimes called, and the actually observed time series. Thus, a time series $z_1, z_2, \ldots, z_N$ of N successive observations is regarded as a sample realization from an infinite population of such time series that could have been generated by the stochastic process. Very often we will omit the word ''stochastic'' from ''stochastic process'' and talk about the ''process.''

1.2.1 Stationary and Nonstationary Stochastic Models for Forecasting and Control

An important class of stochastic models for describing time series, which has received a great deal of attention, comprises what are called *stationary* models. Stationary models assume that the process remains in *statistical equilibrium* with probabilistic properties that do not change over time, in particular varying about a fixed *constant mean level* and with *constant variance*. However, forecasting has been of particular importance in industry, business, and economics, where many time series are often better represented as nonstationary and, in particular, as having no natural constant mean level over time. It is not surprising, therefore, that many of the economic forecasting methods originally proposed by Holt (1957, 1963), Winters (1960), Brown (1962), and the ICI monographs (Coutie, 1964) that used exponentially weighted moving averages can be shown to be appropriate for a particular type of *nonstationary* process. Although such methods are too narrow to deal efficiently with all time series, the fact that they often give the right kind of forecast function supplies a clue to the *kind of nonstationary* model that might be useful in these problems.

The stochastic model for which the exponentially weighted moving average forecast yields minimum mean square error (Muth, 1960) is a member of a class of *nonstationary* processes called autoregressive integrated moving average processes, which are discussed in Chapter 4. This wider class of processes provides a range of models, stationary and nonstationary, that adequately represent many of the time series met in practice. Our approach to forecasting has been first to derive an adequate stochastic model for the particular time series under study. As shown in Chapter 5, once an appropriate model has been determined for the series, the optimal forecasting procedure follows immediately. These forecasting procedures include the exponentially weighted moving average forecast as a special case.

Some Simple Operators. We employ extensively the *backward shift operator* B, which is defined by $Bz_t = z_{t-1}$; hence $B^m z_t = z_{t-m}$. The inverse operation is performed by the *forward shift operator* $F = B^{-1}$ given by $Fz_t = z_{t+1}$; hence $F^m z_t = z_{t+m}$. Another important operator is the *backward difference operator*, ∇, defined by $\nabla z_t = z_t - z_{t-1}$. This can be written in terms of B, since

$$\nabla z_t = z_t - z_{t-1} = (1 - B)z_t$$

Linear Filter Model. The stochastic models we employ are based on an idea originally due to Yule (1927) that an observable time series z_t in which successive values are highly dependent can frequently be regarded as generated from a series of *independent* ''shocks'' a_t. These shocks are *random* drawings from a fixed distribution, usually assumed normal and having mean zero and variance σ_a^2. Such a sequence of independent random variables $a_t, a_{t-1}, a_{t-2}, \ldots$ is called *a white noise* process.

The white noise process a_t is supposed transformed to the process z_t by what is called a *linear filter*, as shown in Figure 1.4. The linear filtering operation simply takes a weighted

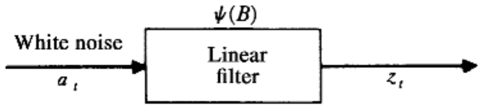

FIGURE 1.4 Representation of a time series as the output from a linear filter.

sum of previous random shocks a_t, so that

$$\begin{aligned} z_t &= \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= \mu + \psi(B) a_t \end{aligned} \tag{1.2.1}$$

In general, μ is a parameter that determines the ''level'' of the process, and

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$$

is the linear operator that transforms a_t into z_t and is called the *transfer function* of the filter. The model representation (1.2.1) can allow for a flexible range of patterns of dependence among values of the process $\{z_t\}$ expressed in terms of the independent (unobservable) random shocks a_t.

The sequence $\psi_1, \psi_2, \ldots$ formed by the weights may, theoretically, be finite or infinite. If this sequence is finite, or infinite and *absolutely summable* in the sense that $\sum_{j=0}^{\infty} |\psi_j| < \infty$, the filter is said to be *stable* and the process z_t is stationary. The parameter μ is then the mean about which the process varies. Otherwise, z_t is nonstationary and μ has no specific meaning except as a reference point for the level of the process.

Autoregressive Models. A stochastic model that can be extremely useful in the representation of certain practically occurring series is the *autoregressive* model. In this model, the current value of the process is expressed as a finite, linear aggregate of *previous values of the process* and a random shock a_t. Let us denote the values of a process at equally spaced times $t,\ t-1,\ t-2,\ \ldots$ by $z_t,\ z_{t-1},\ z_{t-2}, \ldots$. Also let $\tilde{z}_t = z_t - \mu$ be the series of deviations from μ. Then

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t \tag{1.2.2}$$

is called an *autoregressive* (AR) *process of order p*. The reason for this name is that a linear model

$$\tilde{z} = \phi_1 \tilde{x}_1 + \phi_2 \tilde{x}_2 + \cdots + \phi_p \tilde{x}_p + a$$

relating a ''dependent'' variable z to a set of ''independent'' variables $x_1, x_2, \ldots, x_p$, plus a random error term a, is referred to as a *regression* model, and z is said to be ''regressed'' on $x_1, x_2, \ldots, x_p$. In (1.2.2) the variable z is regressed on previous values of itself; hence the model is *auto*regressive. If we define an *autoregressive operator* of order p in terms of the backward shift operator B by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

the autoregressive model (1.2.2) may be written economically as

$$\phi(B)\tilde{z}_t = a_t$$

The model contains $p + 2$ unknown parameters μ, $\phi_1, \phi_2, \ldots, \phi_p, \sigma_a^2$, which in practice have to be estimated from the data. The additional parameter σ_a^2 is the variance of the white noise process a_t.

It is not difficult to see that the autoregressive model is a special case of the linear filter model of (1.2.1). For example, we can eliminate $\tilde{z}_{t-1}$ from the right-hand side of (1.2.2) by substituting

$$\tilde{z}_{t-1} = \phi_1 \tilde{z}_{t-2} + \phi_2 \tilde{z}_{t-3} + \cdots + \phi_p \tilde{z}_{t-p-1} + a_{t-1}$$

Similarly, we can substitute for $\tilde{z}_{t-2}$, and so on, to yield eventually an infinite series in the a's. Consider, specifically, the simple first-order ($p = 1$) AR process, $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$. After m successive substitutions of $\tilde{z}_{t-j} = \phi \tilde{z}_{t-j-1} + a_{t-j}$, $j = 1, \ldots, m$ in the right-hand side we obtain

$$\tilde{z}_t = \phi^{m+1} \tilde{z}_{t-m-1} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots + \phi^m a_{t-m}$$

In the limit as $m \to \infty$ this leads to the convergent infinite series representation $\tilde{z}_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$ with $\psi_j = \phi^j$, $j \geq 1$, provided that $|\phi| < 1$. Symbolically, in the general AR case we have that

$$\phi(B)\tilde{z}_t = a_t$$

is equivalent to

$$\tilde{z}_t = \phi^{-1}(B)a_t = \psi(B)a_t$$

with $\psi(B) = \phi^{-1}(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

Autoregressive processes can be stationary or nonstationary. For the process to be stationary, the ϕ's must be such that the weights $\psi_1, \psi_2, \ldots$ in $\psi(B) = \phi^{-1}(B)$ form a convergent series. The necessary requirement for stationarity is that the autoregressive operator, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, considered as a polynomial in B of degree p, must have all roots of $\phi(B) = 0$ greater than 1 in absolute value; that is, all roots must lie outside the unit circle. For the first-order AR process $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$ this condition reduces to the requirement that $|\phi| < 1$, as the argument above has already indicated.

Moving Average Models. The autoregressive model (1.2.2) expresses the deviation $\tilde{z}_t$ of the process as a *finite* weighted sum of p previous deviations $\tilde{z}_{t-1}, \tilde{z}_{t-2}, \ldots, \tilde{z}_{t-p}$ of the process, plus a random shock a_t. Equivalently, as we have just seen, it expresses $\tilde{z}_t$ as an *infinite* weighted sum of the a's.

Another kind of model, of great practical importance in the representation of observed time series, is the finite *moving average* process. Here we take $\tilde{z}_t$, linearly dependent on a *finite* number q of previous a's. Thus,

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{1.2.3}$$

is called a *moving average* (MA) *process of order q*. The name ''moving average'' is somewhat misleading because the weights $1, -\theta_1, -\theta_2, \ldots, -\theta_q$, which multiply the a's, need not total unity nor need they be positive. However, this nomenclature is in common use, and therefore we employ it.

If we define a *moving average operator* of order q by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

the moving average model may be written economically as

$$\tilde{z}_t = \theta(B) a_t$$

It contains $q + 2$ unknown parameters μ, $\theta_1, \ldots, \theta_q$, σ_a^2, which in practice have to be estimated from the data.

Mixed Autoregressive–Moving Average Models. To achieve greater flexibility in fitting of actual time series, it is sometimes advantageous to include both autoregressive and moving average terms in the model. This leads to the mixed *autoregressive–moving average* (ARMA) model:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{1.2.4}$$

or

$$\phi(B)\tilde{z}_t = \theta(B) a_t$$

The model employs $p + q + 2$ unknown parameters μ, $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, σ_a^2, that are estimated from the data. This model may also be written in the form of the linear filter (1.2.1) as $\tilde{z}_t = \phi^{-1}(B)\theta(B) a_t = \psi(B) a_t$, with $\psi(B) = \phi^{-1}(B)\theta(B)$. In practice, it is frequently true that an adequate representation of actually occurring stationary time series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than 2 and often less than 2. We discuss the classes of autoregressive, moving average, and mixed models in much greater detail in Chapters 3 and 4.

Nonstationary Models. Many series actually encountered in industry or business (e.g., stock prices and sales figures) exhibit nonstationary behavior and in particular do not vary about a fixed mean. Such series may nevertheless exhibit homogeneous behavior over time of a kind. In particular, although the general level about which fluctuations are occurring may be different at different times, the broad behavior of the series, when differences in level are allowed for, may be similar over time. We show in Chapter 4 and later chapters that such behavior may often be represented by a model in terms of a generalized autoregressive operator $\varphi(B)$, in which one or more of the zeros of the polynomial $\varphi(B)$ [i.e., one or more of the roots of the equation $\varphi(B) = 0$] lie on the unit circle. In particular, if there are d unit roots and all other roots lie outside the unit circle, the operator $\varphi(B)$ can be written

$$\varphi(B) = \phi(B)(1 - B)^d$$

where $\phi(B)$ is a stationary autoregressive operator. Thus, a model that can represent homogeneous nonstationary behavior is of the form

$$\varphi(B) z_t = \phi(B)(1 - B)^d z_t = \theta(B) a_t$$

that is,

$$\phi(B) w_t = \theta(B) a_t \tag{1.2.5}$$

where

$$w_t = (1 - B)^d z_t = \nabla^d z_t \tag{1.2.6}$$

Thus, homogeneous nonstationary behavior can sometimes be represented by a model that calls for the dth difference of the process to be stationary. In practice, d is usually 0, 1, or at most 2, with $d = 0$ corresponding to stationary behavior.

The process defined by (1.2.5) and (1.2.6) provides a powerful model for describing stationary and nonstationary time series and is called an *autoregressive integrated moving average process*, of order (p, d, q), or ARIMA(p, d, q) process. The process is defined by

$$w_t = \phi_1 w_{t-1} + \cdots \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{1.2.7}$$

with $w_t = \nabla^d z_t$. Note that if we replace w_t, by $z_t - \mu$, when $d = 0$, the model (1.2.7) includes the stationary mixed model (1.2.4), as a special case, and also the pure autoregressive model (1.2.2) and the pure moving average model (1.2.3).

The reason for inclusion of the word ''integrated'' (which should perhaps more appropriately be ''summed'') in the ARIMA title is as follows. The relationship, which is the inverse to (1.2.6), is $z_t = S^d w_t$, where $S = \nabla^{-1} = (1 - B)^{-1} = 1 + B + B^2 + \cdots$ is the *summation operator* defined by

$$Sw_t = \sum_{j=0}^{\infty} w_{t-j} = w_t + w_{t-1} + w_{t-2} + \cdots$$

Thus, the general ARIMA process may be generated by summing or ''integrating'' the stationary ARMA process $w_t d$ times. In Chapter 9, we describe how a special form of the model (1.2.7) can be employed to represent seasonal time series. The chapter also includes a discussion of regression models where the errors are autocorrelated and follow an ARMA process.

Chapter 10 includes material that may be considered more specialized and that either supplements or extends the material presented in the earlier chapters. The chapter begins with a discussion of unit root testing that may be used as a supplementary tool to determine if a time series is nonstationary and can be made stationary through differencing. This is followed by a discussion of conditionally heteroscedastic models such as the ARCH and GARCH models. These models assume that the conditional variance of an observation given its past vary over time and are useful for modeling time varying volatility in economic and financial time series, in particular. In Chapter 10, we also discuss nonlinear time series models and fractionally integrated long-memory processes that allow for certain more general features in a time series than are possible using the linear ARIMA models.

1.2.2 Transfer Function Models

An important type of dynamic relationship between a continuous input and a continuous output, for which many physical examples can be found, is that in which the deviations of input X and output Y, from appropriate mean values, are related by a *linear* differential equation. In a similar way, for discrete data, in Chapter 11 we represent the transfer relationship between an output Y and an input X, each measured at equispaced times, by the difference equation

$$(1 + \xi_1 \nabla + \cdots + \xi_r \nabla^r) Y_t = (\eta_0 + \eta_1 \nabla + \cdots + \eta_s \nabla^s) X_{t-b} \tag{1.2.8}$$

in which the differential operator $D = d/dt$ is replaced by the difference operator $\nabla = 1 - B$. An expression of the form (1.2.8), containing only a few parameters ($r \le 2, s \le 2$), may often be used as an approximation to a dynamic relationship whose true nature is more complex.

The linear model (1.2.8) may be written equivalently in terms of past values of the input and output by substituting $B = 1 - \nabla$ in (1.2.8), that is,

$$\begin{aligned}(1 - \delta_1 B - \cdots - \delta_r B^r) Y_t &= (\omega_0 - \omega_1 B - \cdots - \omega_s B^s) X_{t-b} \\ &= (\omega_0 B^b - \omega_1 B^{b+1} - \cdots - \omega_s B^{b+s}) X_t \end{aligned} \tag{1.2.9}$$

or

$$\delta(B) Y_t = \omega(B) B^b X_t = \Omega(B) X_t$$

Alternatively, we can say that the output Y_t and the input X_t are linked by a linear filter

$$\begin{aligned} Y_t &= \upsilon_0 X_t + \upsilon_1 X_{t-1} + \upsilon_2 X_{t-2} + \cdots \\ &= \upsilon(B) X_t \end{aligned} \tag{1.2.10}$$

for which the transfer function

$$\upsilon(B) = \upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots \tag{1.2.11}$$

can be expressed as a ratio of two polynomial operators,

$$\upsilon(B) = \frac{\Omega(B)}{\delta(B)} = \delta^{-1}(B)\Omega(B)$$

The linear filter (1.2.10) is said to be *stable* if the series (1.2.11) converges for $|B| \le 1$, equivalently, if the coefficients $\{\upsilon_j\}$ are absolutely summable, $\sum_{j=0}^{\infty} |\upsilon_j| < \infty$. The sequence of weights $\upsilon_0, \upsilon_1, \upsilon_2, \ldots$, which appear in the transfer function (1.2.11), is called the *impulse response function*. We note that for the model (1.2.9), the first b weights $\upsilon_0, \upsilon_1, \ldots, \upsilon_{b-1}$, are zero. A hypothetical impulse response function for the system of Figure 1.2 is shown in the center of that diagram.

Models with Superimposed Noise. We have seen that the problem of estimating an appropriate model, linking an output Y_t and an input X_t, is equivalent to estimating the transfer function $\upsilon(B) = \delta^{-1}(B)\Omega(B)$, for example, specifying the parametric form of the transfer function $\upsilon(B)$ and estimating its parameters. However, this problem is complicated in practice by the presence of noise N_t, which we assume corrupts the true relationship between input and output according to

$$Y_t = \upsilon(B) X_t + N_t$$

where N_t and X_t are independent processes. Suppose, as indicated by Figure 1.5, that the noise N_t can be described by a stationary or nonstationary stochastic model of the form

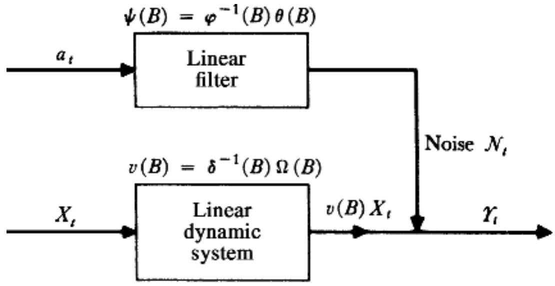

FIGURE 1.5 Transfer function model for dynamic system with superimposed noise model.

(1.2.5) or (1.2.7), that is,

$$N_t = \psi(B)a_t = \varphi^{-1}(B)\theta(B)a_t$$

Then the observed relationship between output and input will be

$$\begin{aligned} Y_t &= v(B)X_t + \psi(B)a_t \\ &= \delta^{-1}(B)\Omega(B)X_t + \varphi^{-1}(B)\theta(B)a_t \end{aligned} \tag{1.2.12}$$

In practice, it is necessary to estimate the transfer function

$$\psi(B) = \varphi^{-1}(B)\theta(B)$$

of the linear filter describing the noise, in addition to the transfer function $v(B) = \delta^{-1}(B)\Omega(B)$, which describes the dynamic relationship between the input and the output. Methods for doing this are discussed in Chapter 12.

1.2.3 Models for Discrete Control Systems

As stated in Section 1.1.5, control is an attempt to compensate for disturbances that infect a system. Some of these disturbances are measurable; others are not measurable and only manifest themselves as unexplained deviations from the target of the characteristic to be controlled. To illustrate the general principles involved, consider the special case where unmeasured disturbances affect the output Y_t of a system, and suppose that feedback control is employed to bring the output as close as possible to the desired target value by adjustments applied to an input variable X_t. This is illustrated in Figure 1.6. Suppose that N_t represents the effect at the output of various unidentified disturbances within the system, which in the absence of control could cause the output to drift away from the desired target value or *set point* T. Then, despite adjustments that have been made to the process, an error

$$\begin{aligned} \varepsilon_t &= Y_t - T \\ &= v(B)X_t + N_t - T \end{aligned}$$

will occur between the output and its target value T. The object is to choose a control equation so that the errors ε have the smallest possible mean square. The control equation expresses the adjustment $x_t = X_t - X_{t-1}$ to be taken at time t, as a function of the present deviation ε_t, previous deviations $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, and previous adjustments $x_{t-1}, x_{t-2}, \ldots$. The mechanism (human, electrical, pneumatic, or electronic) that carries out the control action called for by the control equation is called the *controller*.

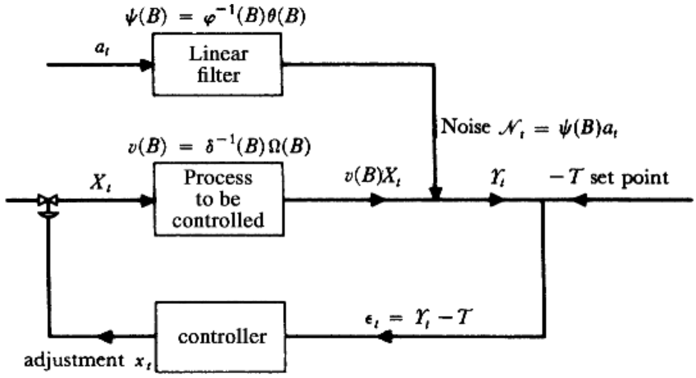

FIGURE 1.6 Feedback control scheme to compensate an unmeasured disturbance N_t.

One procedure for designing a controller is equivalent to forecasting the deviation from target which would occur *if no control were applied*, and then calculating the adjustment that would be necessary to cancel out this deviation. It follows that the forecasting and control problems are closely linked. In particular, if a minimum mean square error forecast is used, the controller will produce minimum mean square error control. To forecast the deviation from target that could occur if no control were applied, it is necessary to build a model

$$N_t = \psi(B)a_t = \varphi^{-1}(B)\theta(B)a_t$$

for the disturbance. Calculation of the adjustment x_t that needs to be applied to the input at time t to cancel out a predicted change at the output requires the building of a dynamic model with transfer function

$$v(B) = \delta^{-1}(B)\Omega(B)$$

which links the input with output. The resulting adjustment x_t will consist, in general, of a linear aggregate of previous adjustments and current and previous control errors. Thus the control equation will be of the form

$$x_t = \zeta_1 x_{t-1} + \zeta_2 x_{t-2} + \cdots + \chi_0 \varepsilon_t + \chi_1 \varepsilon_{t-1} + \chi_2 \varepsilon_{t-2} + \cdots \qquad (1.2.13)$$

where $\zeta_1, \zeta_2, \ldots, \chi_0, \chi_1, \chi_2, \ldots$ are constants.

It turns out that, in practice, minimum mean square error control sometimes results in unacceptably large adjustments x_t to the input variable. Consequently, modified control schemes are employed that restrict the amount of variation in the adjustments. Some of these issues are discussed in Chapter 15.

1.3 BASIC IDEAS IN MODEL BUILDING

1.3.1 Parsimony

We have seen that the mathematical models we need to employ contain certain constants or parameters whose values must be estimated from the data. It is important, in practice, that we employ the *smallest possible* number of parameters for adequate representations. The central role played by this principle of *parsimony* (Tukey, 1961) in the use of parameters will become clearer as we proceed. As a preliminary illustration, we consider the following simple example.

Suppose we fitted a dynamic model (1.2.9) of the form

$$Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s)X_t \tag{1.3.1}$$

when dealing with a system that was adequately represented by

$$(1 - \delta B)Y_t = \omega_0 X_t \tag{1.3.2}$$

The model (1.3.2) contains only two parameters, δ and ω_0, but for s sufficiently large, it could be represented approximately by the model (1.3.1), through

$$Y_t = (1 - \delta B)^{-1}\omega_0 X_t = \omega_0(1 + \delta B + \delta^2 B^2 + \cdots)X_t$$

with $|\delta| < 1$. Because of experimental error, we could easily fail to recognize the relationship between the coefficients in the fitted equation. Thus, we might needlessly fit a relationship like (1.3.1), containing $s + 1$ parameters, where the much simpler form (1.3.2), containing only two, would have been adequate. This could, for example, lead to unnecessarily poor estimation of the output Y_t for given values of the input $X_t, X_{t-1}, \ldots$.

Our objective, then, must be to obtain adequate but parsimonious models. Forecasting and control procedures could be seriously deficient if these models were either inadequate or unnecessarily prodigal in the use of parameters. Care and effort is needed in selecting the model. The process of selection is necessarily iterative; that is, it is a process of evolution, adaptation, or trial and error and is outlined briefly below.

1.3.2 Iterative Stages in the Selection of a Model

If the physical mechanism of a phenomenon were completely understood, it would be possible theoretically to write down a mathematical expression that described it exactly. This would result in a *mechanistic* or *theoretical* model. In most instances the complete knowledge or large experimental resources needed to produce a mechanistic model are not available, and we must resort to an empirical model. Of course, the exact mechanistic model and the exclusively empirical model represent extremes. Models actually employed usually lie somewhere in between. In particular, we may use incomplete theoretical knowledge to indicate a suitable class of mathematical functions, which will then be fitted empirically (e.g., Box and Hunter, 1965); that is, the number of terms needed in the model and the numerical values of the parameters are estimated from experimental data. This is the approach that we adopt in this book. As we have indicated previously, the stochastic and dynamic models we describe can be justified, at least partially, on theoretical grounds as having the right general properties.

It is normally supposed that successive values of the time series under consideration or of the input–output data are available for analysis. If possible, at least 50 and preferably 100 observations or more should be used. In those cases where a past history of 50 or more observations is not available, one proceeds by using experience and past information to derive a preliminary model. This model may be updated from time to time as more data become available.

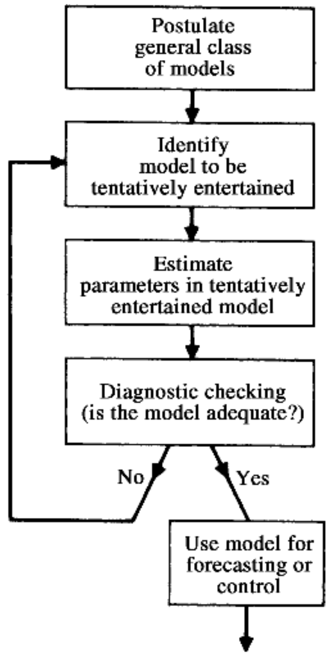

FIGURE 1.7 Stages in the iterative approach to model building.

In fitting dynamic models, a theoretical analysis can sometimes tell us not only the appropriate form for the model, but may also provide us with good estimates of the numerical values of its parameters. These values can then be checked later by analysis of data.

Figure 1.7 summarizes the iterative approach to model building for forecasting and control, which is employed in this book.

1. From the interaction of theory and practice, a *useful class of models* for the purposes at hand is considered.
2. Because this class is too extensive to be conveniently fitted directly to data, rough methods for *identifying* subclasses of these models are developed. Such methods of model identification employ data and knowledge of the system to suggest an appropriate parsimonious subclass of models that may be tentatively entertained. In addition, the identification process can be used to yield rough preliminary estimates of the parameters in the model.
3. The tentatively entertained model is *fitted* to data and its parameters *estimated*. The rough estimates obtained during the identification stage can now be used as starting values in more refined iterative methods for estimating the parameters, such as the nonlinear least squares and maximum likelihood methods.
4. *Diagnostic checks* are applied with the goal of uncovering possible lack of fit and diagnosing the cause. If no lack of fit is indicated, the model is ready to use. If any inadequacy is found, the iterative cycle of identification, estimation, and diagnostic checking is repeated until a suitable representation is found.

Identification, estimation, and diagnostic checking are discussed for univariate time series models in Chapters 6, 7, 8, and 9, for transfer function models in Chapter 12, for intervention models in Chapter 13, and for multivariate time series models in Chapter 14.

The model building procedures will be illustrated using actual time series with numerical calculations performed using the R software and other tools. A brief description of the R software is included in Appendix A1.1 along with references for further study. Exercises at the end of the chapters also make use of the software.

APPENDIX A1.1 USE OF THE R SOFTWARE

The R software for statistical computing and graphics is a common choice for data analysis and development of new statistical methods. R is available as Free Software under the terms of the Free Software Foundations's GNU General Public License in source code form. It compiles and runs on all common operating systems including Windows, MacOS X, and Linux. The main website for the R project is http://www.r-project.org.

The R environment consists of a base system, which is developed and maintained by the R Core Team, and a large set of user contributed packages. The base system provides the source code that implements the basic functionality of R. It also provides a set of standard packages that include commonly used probability distributions, graphical tools, classic datasets from the literature, and a set of statistical methods that include regression analysis and time series analysis. In addition to these base packages, there are now thousands of contributed packages developed by researchers around the world. Packages useful for time series modeling and forecasting include the stats package that is part of the base distribution and several contributed packages that are available for download. These include the TSA package by K-S Chan and Brian Ripley, the astsa package by David Stoffer, the Rmetrics packages fGarch and fUnitRoots for financial time series analysis by Diethelm Wuertz and associates, and the MTS package for multivariate time series analysis by Ruey Tsay. We use many functions from these packages in this book. We also use datasets available for download from the R datasets package, and the TSA and astsa packages.

Both the base system and the contributed packages are distributed through a network of servers called the Comprehensive R Archive Network (CRAN) that can be accessed from the official R website. Contributed packages that are not part of the base distribution can be installed directly from the R prompt ‘‘>’’ using the command install.package(). Under the Windows system, the installation can also be done from a drop-down list. The command will prompt the user to select a *CRAN Mirror*, after which a list of packages available for installation appears. To use a specific package, it also needs to be loaded into the system at the start of each session. For example, the TSA package can be loaded using the commands library(TSA) or require(TSA). The command data() will list all datasets available in the loaded packages. The command data(airquality) will load the dataset airquality from the R datasets package into memory. Data stored in a text file can be read into R using the command is read.table. For a .csv file, the command is read.csv. To get help on specific functions, e.g. the arima function which fits an ARIMA model to a time series, type help(arima) or ?arima.

R is object-oriented software and allows the user to create many objects. For example, the command ts() will create a time series object. This has advantages for plotting the time series and for certain other applications. However, it is not necessary to create a time series object for many of the applications discussed in this book. The structure of the data in R can be examined using commands such as class(), str(), and summary().

The data used for illustration in this book, as well as in some of the exercises, include a set of time series listed in Part Five of the book. These series are also available at http://pages.cs.wisc.edu/ reinsel/bjr-data/index.html. At least three of the series are also included in the R datasets package and can be accessed using the data() command described above. Some of the exercises require the use of R and it will be assumed that the reader is already familiar with the basics of R, which can be obtained by working through relevant chapters of texts such as Crawley (2007) and Adler (2010). Comprehensive documentation in the form of manuals, contributed documents, online help pages, and FAQ sheets is also available on the R website. Since R builds on the S language, a useful reference book is also Venables and Ripley (2002).

EXERCISES

1.1. The dataset airquality in the R datasets package includes information on daily air quality measurements in New York, May to September 1973. The variables included are mean ozone levels at Roosevelt Island, solar radiation at Central Park, average wind speed at LaGuardia Airport, and maximum daily temperature at LaGuardia Airport; see help(airquality) for details.

- **(a)** Load the dataset into R.
- **(b)** Investigate the structure of the dataset.
- **(c)** Plot each of the four series mentioned above using the plot() command in R; see help(plot) for details and examples.
- **(d)** Comment on the behavior of the four series. Do you see any issues that may require special attention in developing a time series model for each of the four series.

1.2. Monthly totals of international airline passengers (in thousands of passengers), January 1949–December 1960, are available as Series G in Part Five of this book. The data are also available as series AirPassengers in the R datasets package.

- **(a)** Load the dataset into R and examine the structure of the data.
- **(b)** Plot the data using R and describe the behavior of the series.
- **(c)** Perform a log transformation of the data and plot the resulting series. Compare the behavior of the original and log-transformed series. Do you see an advantage in using a log transformation for modeling purposes?

1.3. Download a time series of your choosing from the Internet. Note that financial and economic time series are available from sources such as Google Finance and the Federal Reserve Economic Data (FRED) of Federal Reserve Bank in St. Louis, Missouri, while climate data is available from from NOAA's National Climatic Data Center (NCDC).

- **(a)** Store the data in a text file or a .csv file and read the data into R.
- **(b)** Examine the properties of your series using plots or other appropriate tools.
- **(c)** Does your time series appear to be stationary? If not, would differencing and/or some other transformation make the series stationary?