CHAPTER 1

INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING

It is difficult to make predictions, especially about the future

NEILS BOHR, *Danish physicist*

1.1 THE NATURE AND USES OF FORECASTS

A **forecast** is a prediction of some future event or events. As suggested by Neils Bohr, making good predictions is not always easy. Famously "bad" forecasts include the following from the book *Bad Predictions*:

- "The population is constant in size and will remain so right up to the end of mankind." *L'Encyclopedie*, 1756.
- "1930 will be a splendid employment year." U.S. Department of Labor, *New Year's Forecast* in 1929, just before the market crash on October 29.
- "Computers are multiplying at a rapid rate. By the turn of the century there will be 220,000 in the U.S." *Wall Street Journal*, 1966.

Montgomery, Jennings and Kulahci Introduction to Time Series Analysis and Forecasting, Third Edition.
Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci.

Companion Website: www.wiley.com/go/montgomery/timeseriesforecasting3e

Forecasting is an important problem that spans many fields including business and industry, government, economics, environmental sciences, public health, medicine, social science, politics, and finance. Forecasting problems are often classified as short-term, medium-term, and long-term. Short-term forecasting problems involve predicting events only a few time periods (days, weeks, and months) into the future. Medium-term forecasts extend from 1 to 2 years into the future, and long-term forecasting problems can extend beyond that by many years. Short- and medium-term forecasts are required for activities that range from operations management to budgeting and selecting new research and development projects. Long-term forecasts impact issues such as strategic planning. Short- and medium-term forecasting is typically based on identifying, modeling, and extrapolating the patterns found in historical data. Because these historical data usually exhibit inertia and do not change dramatically very quickly, statistical methods are very useful for short- and medium-term forecasting. This book is about the use of these statistical methods.

Most forecasting problems involve the use of time series data. A **time series** is a time-oriented or chronological sequence of observations on a variable of interest. For example, Figure 1.1 shows the market yield on US Treasury Securities at 10-year constant maturity from April 1953 through

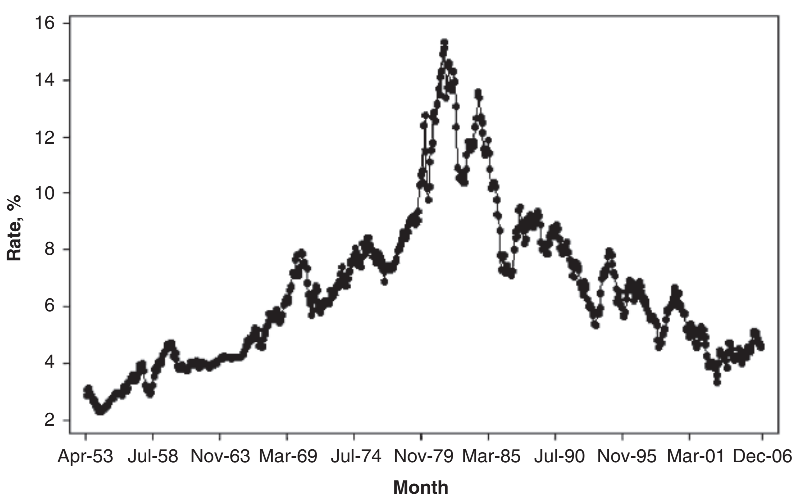

FIGURE 1.1 Time series plot of the market yield on US Treasury Securities at 10-year constant maturity. *Source:* US Treasury.

December 2006 (data in Appendix B, Table B.1). This graph is called a **time series plot**. The rate variable is collected at equally spaced time periods, as is typical in most time series and forecasting applications. Many business applications of forecasting utilize daily, weekly, monthly, quarterly, or annual data, but any reporting interval may be used. Furthermore, the data may be instantaneous, such as the viscosity of a chemical product at the point in time where it is measured; it may be cumulative, such as the total sales of a product during the month; or it may be a statistic that in some way reflects the activity of the variable during the time period, such as the daily closing price of a specific stock on the New York Stock Exchange.

Because time series data exhibits the inertial effects mentioned previously, they usually do not satisfy the usual assumptions made in most statistical methods. That is, time series data are usually not independent. This means that special statistical methods that takes this into account are required for most forecasting and time series analysis problems. Those methods are the focus of this work.

The reason that forecasting is so important is that prediction of future events is a critical input into many types of planning and decision-making processes, with application to areas such as the following:

1. *Operations Management*. Business organizations routinely use forecasts of product sales or demand for services in order to schedule production, control inventories, manage the supply chain, determine staffing requirements, and plan capacity. Forecasts may also be used to determine the mix of products or services to be offered and the locations at which products are to be produced.
2. *Marketing*. Forecasting is important in many marketing decisions. Forecasts of sales response to advertising expenditures, new promotions, or changes in pricing polices enable businesses to evaluate their effectiveness, determine whether goals are being met, and make adjustments.
3. *Finance and Risk Management*. Investors in financial assets are interested in forecasting the returns from their investments. These assets include but are not limited to stocks, bonds, and commodities; other investment decisions can be made relative to forecasts of interest rates, options, and currency exchange rates. Financial risk management requires forecasts of the volatility of asset returns so that the risks associated with investment portfolios

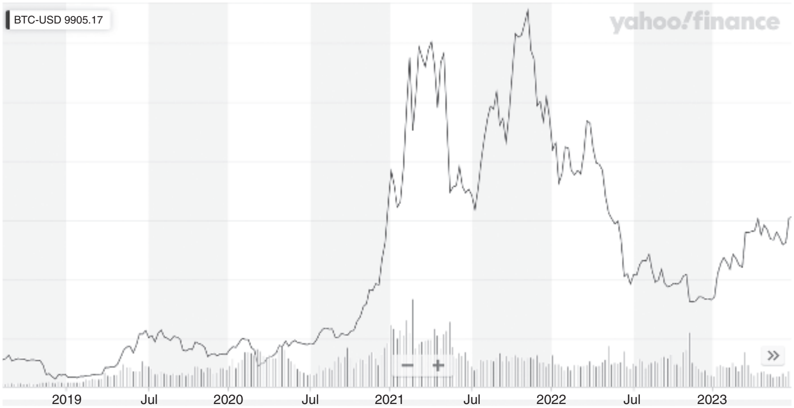

FIGURE 1.2 Five years of Bitcoin prices (from Yahoo Finance).

can be evaluated and insured, and so that financial derivatives can be properly priced.

As an example of financial data, consider the Bitcoin price history for the most recent five years shown in Figure 1.2. Bitcoin is a cryptocurrency introduced in early 2009 although standard pricing did not begin until about a year later. The graph shows that there has been considerable growth in Bitcoin prices, but also considerable volatility. Investors and currency traders would be interested in modeling and forecasting the performance of this asset. However, the inherent volatility in Bitcoin price would make this a very challenging task.

4. *Economics*. Governments, financial institutions, and policy organizations require forecasts of major economic variables, such as gross domestic product, population growth, unemployment, interest rates, inflation, job growth, production, and consumption. These forecasts are an integral part of the guidance behind monetary and fiscal policy, and budgeting plans and decisions made by governments. They are also instrumental in the strategic planning decisions made by business organizations and financial institutions.
5. *Industrial Process Control*. Forecasts of the future values of critical quality characteristics of a production process can help determine when important controllable variables in the process should be changed, or if the process should be shut down and overhauled. Feedback and feedforward control schemes are widely used in

monitoring and adjustment of industrial processes, and predictions of the process output are an integral part of these schemes.

6. *Demography*. Forecasts of population by country and regions are made routinely, often stratified by variables such as gender, age, and race. Demographers also forecast births, deaths, and migration patterns of populations. Governments use these forecasts for planning policy and social service actions, such as spending on health care, retirement programs, and antipoverty programs. Many businesses use forecasts of populations by age groups to make strategic plans regarding developing new product lines or the types of services that will be offered.
7. *Public Health Applications*. As an example of the use of time series analysis in the public health arena, let us consider the recent COVID-19 pandemic. This is also known as the **coronavirus pandemic**, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus was first identified in an outbreak in the China in December 2019. Attempts to contain it there failed, allowing the virus to spread worldwide in 2020. The pandemic triggered social and economic disruption around the world, including a global recession. There were widespread supply shortages, including food shortages, resulting from supply chain disruptions. Mitigation strategies including travel restriction, business and school closures, social distancing measures, masking mandates, testing, contact tracing of infected individuals, and remote working were widespread. COVID-19 vaccines became available in late 2020 and have been widely deployed. As of mid-2023, the pandemic had caused over 700,000,000 cases and approximately 6.9 million deaths.

Figure 1.3 shows a time series plot of daily new cases from early 2020 to mid-2023. Figure 1.4 shows a plot of daily deaths from mid-2020 through mid-2023. The number of deaths declined rapidly over the last year shown in the graph due to the various mitigation strategies and the widespread availability and use of effective vaccines. Public health agencies at the national, state, and local level frequently made data such as this available. There was also interest in hospitalizations arising from the disease, as there was some concerns that hospital resources would be overwhelmed by the number of cases requiring that level of care.

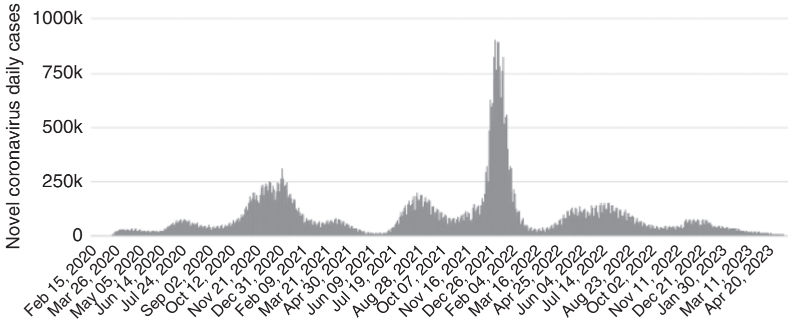

FIGURE 1.3 Daily new cases of COVID-19.

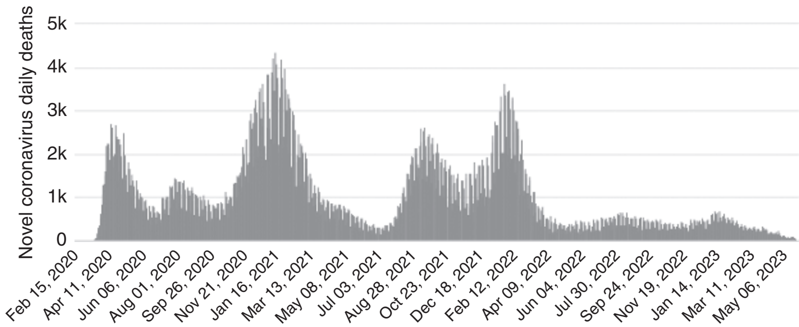

FIGURE 1.4 Daily deaths from COVID-19.

A time series analysist could use this data to predict cases, deaths, and hospitalizations. It could also be possible to include the introduction of the mitigation measures including vaccines in the analysis to determine the potential effectiveness of these measures in reducing the spread of severity of the disease. A technique called intervention analysis can be used to do this. Intervention analysis is discussed in this book in Chapter 6.

These are only a few of the many different situations where forecasts are required in order to make good decisions. Despite the wide range of problem situations that require forecasts, there are only two broad types of forecasting techniques—qualitative methods and quantitative methods.

Qualitative forecasting techniques are often subjective in nature and require judgment on the part of experts. Qualitative forecasts are often used in situations where there is little or no historical data on which to base the forecast. An example would be the introduction of a new product, for which there is no relevant history. In this situation, the company might use the expert opinion of sales and marketing personnel to subjectively estimate product sales during the new product introduction phase of its life cycle. Sometimes qualitative forecasting methods make use of marketing tests, surveys of potential customers, and experience with the sales performance of other products (both their own and those of competitors). However, although some data analysis may be performed, the basis of the forecast is subjective judgment.

Perhaps the most formal and widely known qualitative forecasting technique is the **Delphi Method**. This technique was developed by the RAND Corporation (see Dalkey [1967]). It employs a panel of experts who are assumed to be knowledgeable about the problem. The panel members are physically separated to avoid their deliberations being impacted either by social pressures or by a single dominant individual. Each panel member responds to a questionnaire containing a series of questions and returns the information to a coordinator. Following the first questionnaire, subsequent questions are submitted to the panelists along with information about the opinions of the panel as a group. This allows panelists to review their predictions relative to the opinions of the entire group. After several rounds, it is hoped that the opinions of the panelists converge to a consensus, although achieving a consensus is not required and justified differences of opinion can be included in the outcome. Qualitative forecasting methods are not emphasized in this book.

Quantitative forecasting techniques make formal use of historical data and a **forecasting model**. The model formally summarizes patterns in the data and expresses a statistical relationship between previous and current values of the variable. Then the model is used to project the patterns in the data into the future. In other words, the forecasting model is used to extrapolate past and current behavior into the future. There are several types of forecasting models in general use. The three most widely used are regression models, smoothing models, and general time series models. Regression models make use of relationships between the variable of interest and one or more related predictor variables. Sometimes regression models are called **causal forecasting models,** because the predictor variables are assumed to describe the forces that cause or drive the observed values of the variable of interest. An example would be using data on house purchases as a predictor variable to forecast furniture sales.

The method of least squares is the formal basis of most regression models. **Smoothing models** typically employ a simple function of previous observations to provide a forecast of the variable of interest. These methods may have a formal statistical basis, but they are often used and justified heuristically on the basis that they are easy to use and produce satisfactory results. General **time series models** employ the statistical properties of the historical data to specify a formal model and then estimate the unknown parameters of this model (usually) by least squares. In subsequent chapters, we will discuss all three types of quantitative forecasting models.

The form of the forecast can be important. We typically think of a forecast as a single number that represents our best estimate of the future value of the variable of interest. Statisticians would call this a **point estimate** or **point forecast.** Now these forecasts are almost always wrong; that is, we experience **forecast error**. Consequently, it is usually a good practice to accompany a forecast with an estimate of how large a forecast error might be experienced. One way to do this is to provide a **prediction interval** (PI) to accompany the point forecast. The PI is a range of values for the future observation, and it is likely to prove far more useful in decision-making than a single number. We will show how to obtain PIs for most of the forecasting methods discussed in the book.

Other important features of the forecasting problem are the **forecast horizon** and the **forecast interval.** The forecast horizon is the number of future periods for which forecasts must be produced. The horizon is often dictated by the nature of the problem. For example, in production planning, forecasts of product demand may be made on a monthly basis. Because of the time required to change or modify a production schedule, ensure that sufficient raw material and component parts are available from the supply chain, and plan the delivery of completed goods to customers or inventory facilities, it would be necessary to forecast up to 3 months ahead. The forecast horizon is also often called the forecast **lead time.** The **forecast interval** is the frequency with which new forecasts are prepared. For example, in production planning, we might forecast demand on a monthly basis, for up to 3 months in the future (the lead time or horizon), and prepare a new forecast each month. Thus the forecast interval is 1 month, the same as the basic period of time for which each forecast is made. If the forecast lead time is always the same length, say, T periods, and the forecast is revised each time period, then we are employing a **rolling** or **moving horizon** forecasting approach.

This system updates or revises the forecasts for $T-1$ of the periods in the horizon and computes a forecast for the newest period T. This rolling horizon approach to forecasting is widely used when the lead time is several periods long.

1.2 SOME EXAMPLES OF TIME SERIES

Time series plots can reveal **patterns** such as random, trends, level shifts, periods or cycles, unusual observations, or a combination of patterns. Patterns commonly found in time series data are discussed next with examples of situations that drive the patterns.

The sales of a mature pharmaceutical product may remain relatively flat in the absence of unchanged marketing or manufacturing strategies. Weekly sales of a generic pharmaceutical product shown in Figure 1.5 appear to be constant over time, at about $10{,}400 \times 10^3$ units, in a random sequence with no obvious patterns (data in Appendix B, Table B.2).

To assure conformance with customer requirements and product specifications, the production of chemicals is monitored by many characteristics. These may be input variables such as temperature and flow rate, and output properties such as viscosity and purity.

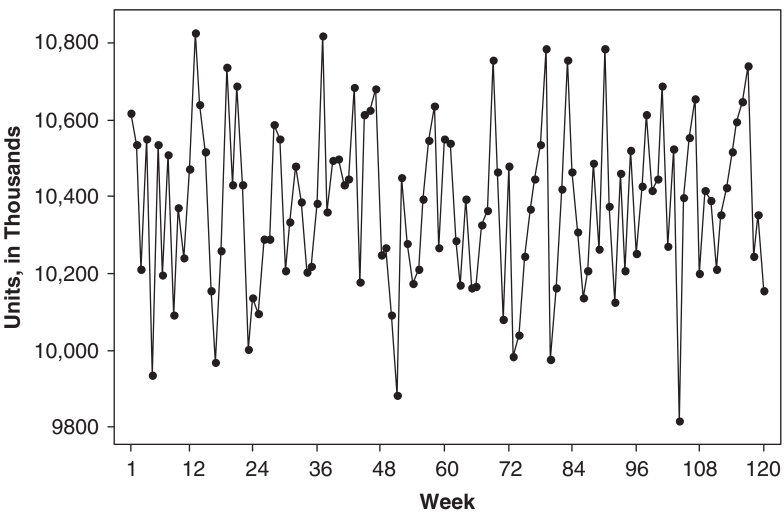

FIGURE 1.5 Pharmaceutical product sales.

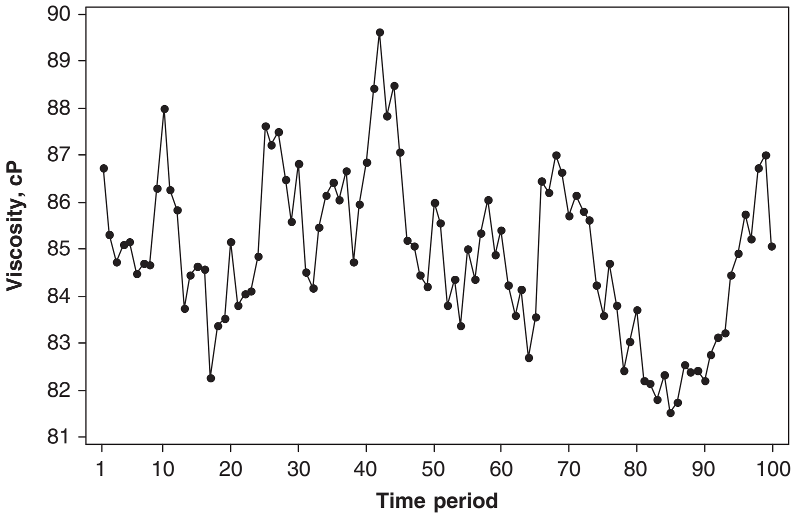

FIGURE 1.6 Chemical process viscosity readings.

Due to the continuous nature of chemical manufacturing processes, output properties often are **positively autocorrelated;** that is, a value above the long-run average tends to be followed by other values above the average, while a value below the average tends to be followed by other values below the average.

The viscosity readings plotted in Figure 1.6 exhibit autocorrelated behavior, tending to a long-run average of about 85 centipoises (cP), but with a structured, not completely random, appearance (data in Appendix B, Table B.3). Some methods for describing and analyzing autocorrelated data will be described in Chapter 2.

The USDA National Agricultural Statistics Service publishes agricultural statistics for many commodities, including the annual production of dairy products such as butter, cheese, ice cream, milk, yogurt, and whey. These statistics are used for market analysis and intelligence, economic indicators, and identification of emerging issues.

Blue and gorgonzola cheese is one of 32 categories of cheese for which data are published. The annual US production of blue and gorgonzola cheeses (in 10^3 lb) is shown in Figure 1.7 (data in Appendix B, Table B.4). Production quadrupled from 1950 to 1997, and the **linear trend** has a constant positive slope with random, year-to-year variation.

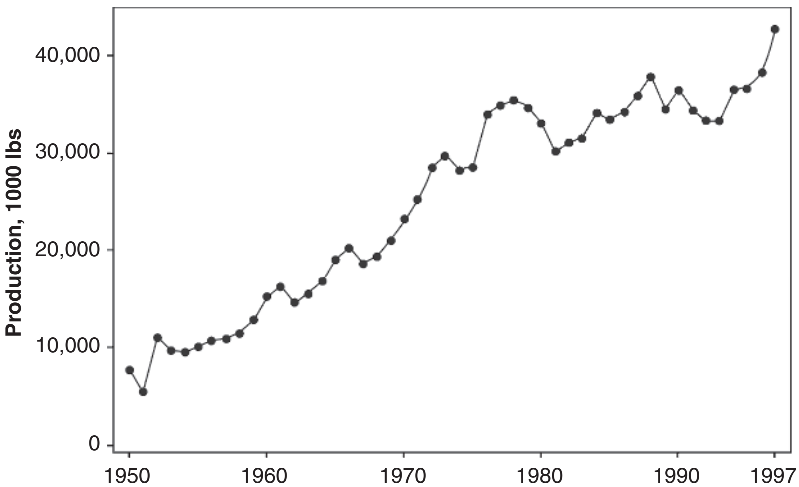

FIGURE 1.7 The US annual production of blue and gorgonzola cheeses. *Source:* USDA–NASS.

The US Census Bureau publishes historic statistics on manufacturers' shipments, inventories, and orders. The statistics are based on North American Industry Classification System (NAICS) code and are utilized for purposes such as measuring productivity and analyzing relationships between employment and manufacturing output.

The manufacture of beverage and tobacco products is reported as part of the nondurable subsector. The plot of monthly beverage product shipments (Figure 1.8) reveals an overall increasing trend, with a distinct **cyclic pattern** that is repeated within each year. January shipments appear to be the lowest, with highs in May and June (data in Appendix B, Table B.5). This monthly, or **seasonal,** variation may be attributable to some cause such as the impact of weather on the demand for beverages. Techniques for making seasonal adjustments to data in order to better understand general trends will be discussed in Chapter 2.

To determine whether the Earth is warming or cooling, scientists look at annual mean temperatures. At a single station, the warmest and the coolest temperatures in a day are averaged. Averages are then calculated at stations all over the Earth, over an entire year. The change in global annual mean surface air temperature is calculated from a base established from 1951 to 1980, and the result is reported as an "anomaly."

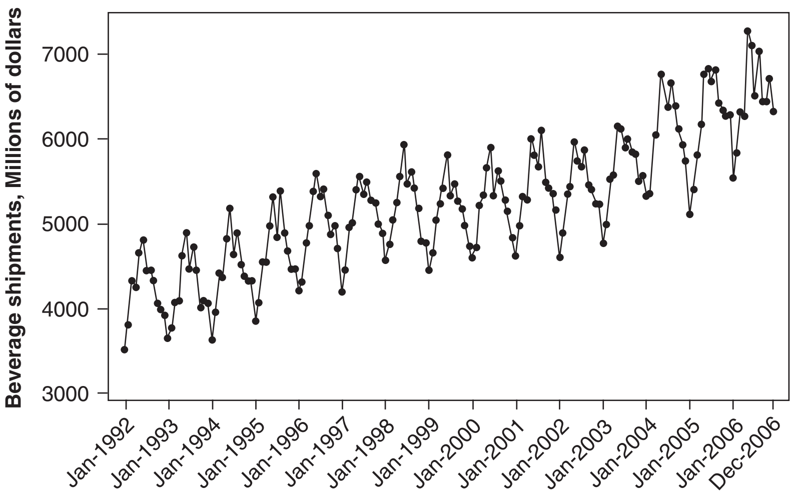

FIGURE 1.8 The US beverage manufacturer monthly product shipments, unadjusted. *Source:* US Census Bureau.

The plot of the annual mean anomaly in global surface air temperature (Figure 1.9) shows an increasing trend since 1880; however, the slope, or rate of change, varies with time periods (data in Appendix B, Table B.6). While the slope in earlier time periods appears to be constant, slightly increasing, or slightly decreasing, the slope from about 1975 to the present appears much steeper than the rest of the plot.

Business data such as stock prices and interest rates often exhibit **nonstationary** behavior; that is, the time series has no natural mean. The daily closing price adjusted for stock splits of Whole Foods Market (WFMI) stock in 2001 (Figure 1.10) exhibits a combination of patterns for both mean level and slope (data in Appendix B, Table B.7).

While the price is constant in some short time periods, there is no consistent mean level over time. In other time periods, the price changes at different rates, including occasional abrupt shifts in level. This is an example of nonstationary behavior, which will be discussed in Chapter 2.

The Current Population Survey (CPS) or "household survey" prepared by the US Department of Labor, Bureau of Labor Statistics, contains national data on employment, unemployment, earnings, and other labor market topics by demographic characteristics. The data are used to report on the employment situation, for projections with impact on hiring and training, and for a multitude of other business planning activities. The data

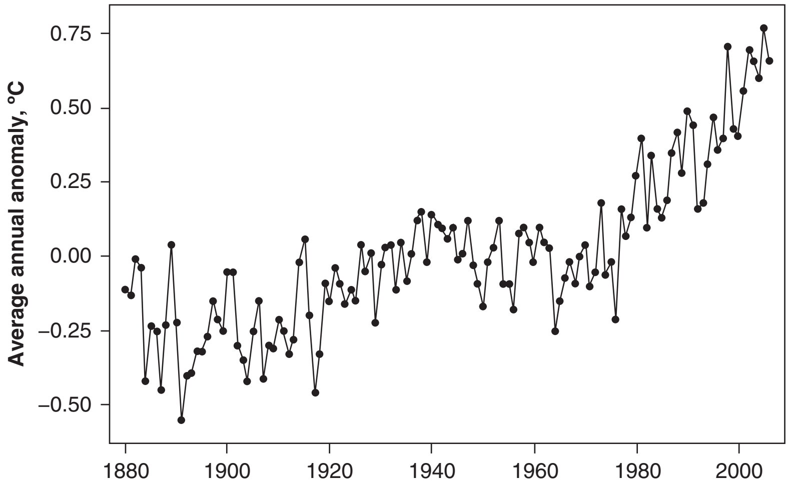

FIGURE 1.9 Global mean surface air temperature annual anomaly. *Source:* NASA-GISS.

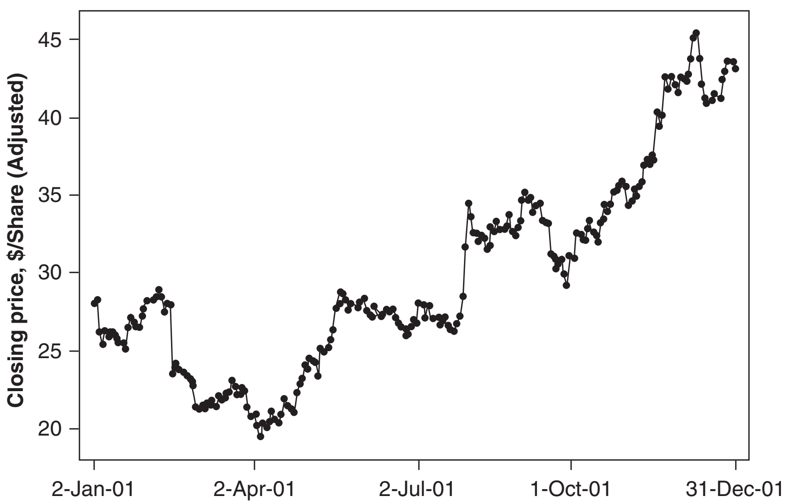

FIGURE 1.10 Whole foods market stock price, daily closing adjusted for splits.

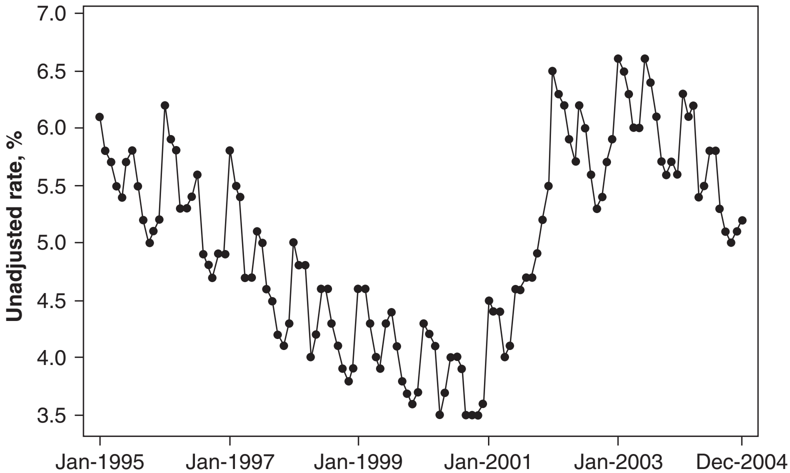

FIGURE 1.11 Monthly unemployment rate—full-time labor force, unadjusted. *Source:* US Department of Labor-BLS.

are reported unadjusted and with seasonal adjustment to remove the effect of regular patterns that occur each year.

The plot of monthly unadjusted unemployment rates (Figure 1.11) exhibits a mixture of patterns, similar to Figure 1.8 (data in Appendix B, Table B.8). There is a distinct cyclic pattern within a year; January, February, and March generally have the highest unemployment rates. The overall level is also changing, from a gradual decrease, to a steep increase, followed by a gradual decrease. The use of seasonal adjustments as described in Chapter 2 makes it easier to observe the nonseasonal movements in time series data.

Solar activity has long been recognized as a significant source of noise impacting consumer and military communications, including satellites, cell phone towers, and electric power grids. The ability to accurately forecast solar activity is critical to a variety of fields. The International Sunspot Number *R* is the oldest solar activity index. The number incorporates both the number of observed sunspots and the number of observed sunspot groups. In Figure 1.12, the plot of annual sunspot numbers reveals cyclic patterns of varying magnitudes (data in Appendix B, Table B.9).

In addition to assisting in the identification of steady-state patterns, time series plots may also draw attention to the occurrence of **atypical events.** Weekly sales of a generic pharmaceutical product dropped due to limited availability resulting from a fire at one of the four production

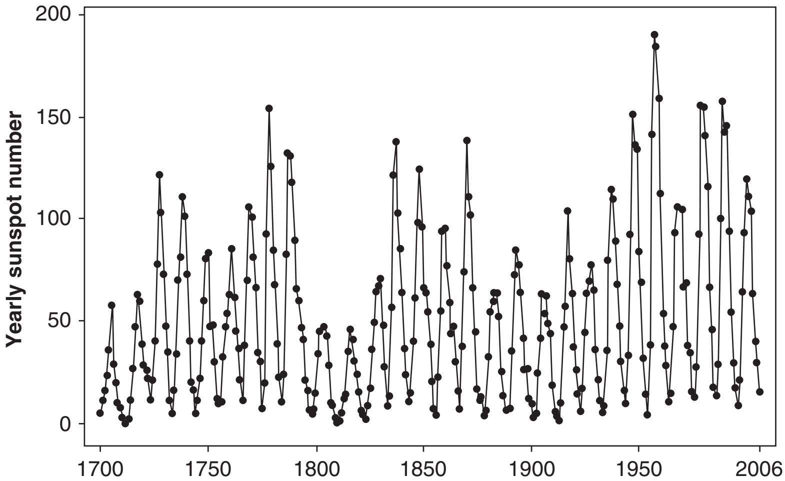

FIGURE 1.12 The international sunspot number. *Source:* SIDC.

facilities. The 5-week reduction is apparent in the time series plot of weekly sales shown in Figure 1.13.

Another type of unusual event may be the failure of the data measurement or collection system. After recording a vastly different viscosity

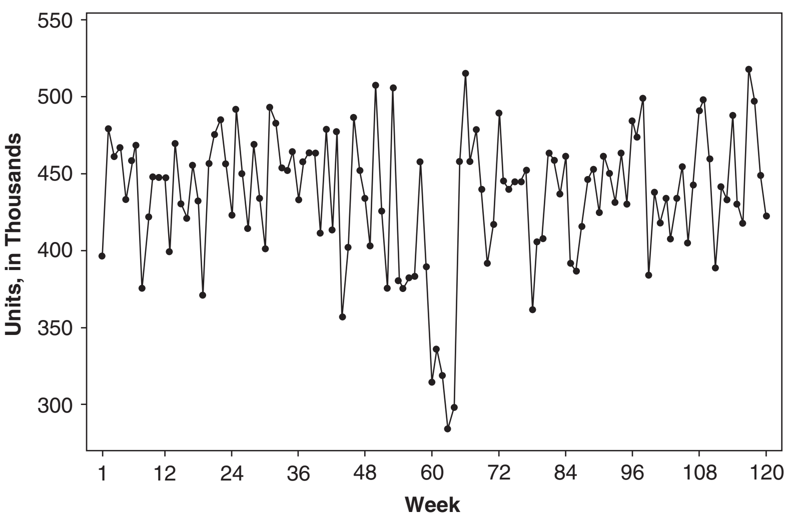

FIGURE 1.13 Pharmaceutical product sales.

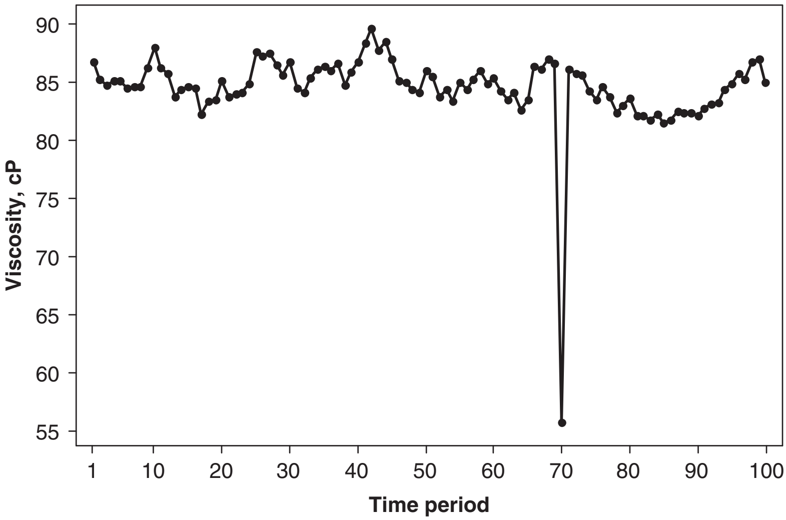

FIGURE 1.14 Chemical process viscosity readings, with sensor malfunction.

reading at time period 70 (Figure 1.14), the measurement system was checked with a standard and determined to be out of calibration. The cause was determined to be a malfunctioning sensor.

1.3 THE FORECASTING PROCESS

A process is a series of connected activities that transform one or more inputs into one or more outputs. All work activities are performed in processes, and forecasting is no exception. The activities in the forecasting process are:

1. Problem definition
2. Data collection
3. Data analysis
4. Model selection and fitting
5. Model validation
6. Forecasting model deployment
7. Monitoring forecasting model performance

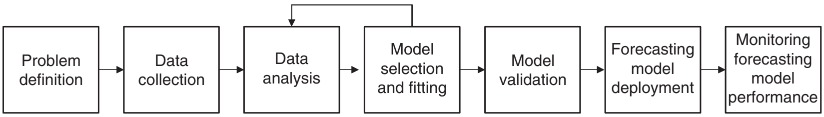

FIGURE 1.15 The forecasting process.

These activities are shown in Figure 1.15.

Problem definition involves developing understanding of how the forecast will be used along with the expectations of the "customer" (the user of the forecast). Questions that must be addressed during this phase include the desired form of the forecast (e.g., are monthly forecasts required), the forecast horizon or lead time, how often the forecasts need to be revised (the forecast interval), and what level of forecast accuracy is required in order to make good business decisions. This is also an opportunity to introduce the decision makers to the use of prediction intervals as a measure of the risk associated with forecasts, if they are unfamiliar with this approach. Often it is necessary to go deeply into many aspects of the business system that requires the forecast to properly define the forecasting component of the entire problem. For example, in designing a forecasting system for inventory control, information may be required on issues such as product shelf life or other aging considerations, the time required to manufacture or otherwise obtain the products (production lead time), and the economic consequences of having too many or too few units of product available to meet customer demand. When multiple products are involved, the level of aggregation of the forecast (e.g., do we forecast individual products or families consisting of several similar products) can be an important consideration. Much of the ultimate success of the forecasting model in meeting the customer expectations is determined in the problem definition phase.

Data collection consists of obtaining the relevant history for the variable(s) that are to be forecast, including historical information on potential predictor variables.

The key here is "relevant"; often information collection and storage methods and systems change over time and not all historical data are useful for the current problem. Often it is necessary to deal with missing values of some variables, potential outliers, or other data-related problems that have occurred in the past. During this phase, it is also useful to begin planning how the data collection and storage issues in the future will be handled so that the reliability and integrity of the data will be preserved.

Data analysis is an important preliminary step to the selection of the forecasting model to be used. Time series plots of the data should be constructed and visually inspected for recognizable patterns, such as trends and seasonal or other cyclical components. A trend is evolutionary movement, either upward or downward, in the value of the variable. Trends may be long-term or more dynamic and of relatively short duration. Seasonality is the component of time series behavior that repeats on a regular basis, such as each year. Sometimes we will smooth the data to make identification of the patterns more obvious (data smoothing will be discussed in Chapter 2). Numerical summaries of the data, such as the sample mean, standard deviation, percentiles, and autocorrelations, should also be computed and evaluated. Chapter 2 will provide the necessary background to do this. If potential predictor variables are available, scatter plots of each pair of variables should be examined. Unusual data points or potential **outliers** should be identified and flagged for possible further study. The purpose of this preliminary data analysis is to obtain some "feel" for the data, and a sense of how strong the underlying patterns such as trend and seasonality are. This information will usually suggest the initial types of quantitative forecasting methods and models to explore.

Model selection and fitting consists of choosing one or more forecasting models and fitting the model to the data. **By fitting**, we mean estimating the unknown model parameters, usually by the method of least squares. In subsequent chapters, we will present several types of time series models and discuss the procedures of model fitting. We will also discuss methods for evaluating the quality of the model fit, and determining if any of the underlying assumptions have been violated. This will be useful in discriminating between different candidate models.

Model validation consists of an evaluation of the forecasting model to determine how it is likely to perform in the intended application. This must go beyond just evaluating the "fit" of the model to the historical data and must examine what magnitude of forecast errors will be experienced when the model is used to forecast "fresh" or new data. The fitting errors will always be smaller than the forecast errors, and this is an important concept that we will emphasize in this book. A widely used method for validating a forecasting model before it is turned over to the customer is to employ some form of **data splitting,** where the data are divided into two segments—a fitting segment and a forecasting segment. The model is fit to only the fitting data segment, and then forecasts from that model are simulated for the observations in the forecasting segment. This can provide useful guidance on how the forecasting model will perform when exposed to new data and can be a valuable approach for discriminating between competing forecasting models.

Forecasting model deployment involves getting the model and the resulting forecasts in use by the customer. It is important to ensure that the customer understands how to use the model and that generating timely forecasts from the model becomes as routine as possible. Model maintenance, including making sure that data sources and other required information will continue to be available to the customer is also an important issue that impacts the timeliness and ultimate usefulness of forecasts.

Monitoring forecasting model performance should be an ongoing activity after the model has been deployed to ensure that it is still performing satisfactorily. It is the nature of forecasting that conditions change over time, and a model that performed well in the past may deteriorate in performance. Usually performance deterioration will result in larger or more systematic forecast errors. Therefore monitoring of forecast errors is an essential part of good forecasting system design. **Control charts** of forecast errors are a simple but effective way to routinely monitor the performance of a forecasting model. We will illustrate approaches to monitoring forecast errors in subsequent chapters.

1.4 DATA FOR FORECASTING

1.4.1 The Data Warehouse

Developing time series models and using them for forecasting requires data on the variables of interest to decision-makers. The data are the raw materials for the modeling and forecasting process. The terms **data** and **information** are often used interchangeably, but we prefer to use the term data as that seems to reflect a more raw or original form, whereas we think of information as something that is extracted or synthesized from data. The output of a forecasting system could be thought of as information, and that output uses data as an input.

In most modern organizations data regarding sales, transactions, company financial and business performance, supplier performance, and customer activity and relations are stored in a repository known as a **data warehouse**. Sometimes this is a single data storage system; but as the volume of data handled by modern organizations grows rapidly, the data warehouse has become an integrated system comprised of components that are physically and often geographically distributed, such as cloud data storage. The data warehouse must be able to organize, manipulate, and integrate data from multiple sources and different organizational information systems. The basic functionality required includes data extraction, data transformation, and data loading. Data extraction refers to obtaining data from internal sources and from external sources such as third party vendors or government entities and financial service organizations. Once the data are extracted, the transformation stage involves applying rules to prevent duplication of records and dealing with problems such as missing information. Sometimes we refer to the transformation activities as **data wrangling and cleaning**. We will discuss some of these operations subsequently. Finally, the data are loaded into the data warehouse where they are available for modeling and analysis.

Data quality has several dimensions. Five important ones that have been described in the literature are accuracy, timeliness, completeness, representativeness, and consistency. Accuracy is probably the oldest dimension of data quality and refers to how close that data conform to its “real” values. Real values are alternative sources that can be used for verification purposes. For example, do sales records match payments to accounts receivable records (although the financial records may occur in later time periods because of payment terms and conditions, discounts, etc.)? Timeliness means that the data are as current as possible. Infrequent updating of data can seriously impact developing a time series model that is going to be used for relatively short-term forecasting. In many time series model applications the time between the occurrence of the real-world event and its entry into the data warehouse must be as short as possible to facilitate model development and use. Completeness means that the data content is complete, with no missing data and no outliers. As an example of representativeness, suppose that the end use of the time series model is to forecast customer demand for a product or service, but the organization only records booked orders and the date of fulfillment. This may not accurately reflect demand, because the orders can be booked before the desired delivery period and the date of fulfillment can take place in a different period than the one required by the customer. Furthermore, orders that are lost because of product unavailability or unsatisfactory delivery performance are not recorded. In these situations demand can differ dramatically from sales. Data cleaning methods can often be used to deal with some problems of completeness. Consistency refers to how closely data records agree over time in format, content, meaning, and structure. In many organizations how data are collected and stored evolves over time; definitions change and even the types of data that are collected change. For example, consider monthly data. Some organizations define "months" that coincide with the traditional calendar definition. But because months have different numbers of days that can induce patterns in monthly data, some organizations prefer to define a year as consisting of 13 "months" each consisting of 4 weeks.

It has been suggested that the output data that reside in the data warehouse are similar to the output of a manufacturing process, where the raw data are the input. Just as in manufacturing and other service processes, the data production process can benefit by the application of quality management and control tools. Jones-Farmer et al. (2014) describe how statistical quality control methods, specifically control charts, can be used to enhance data quality in the data production process.

1.4.2 Data Wrangling and Cleaning

Data wrangling and data cleaning are two important steps that often need to be performed in advance of time series analysis, model, and forecasting. **Data wrangling** is the process of transforming and mapping data from one format into another the purpose of making it more suitable for a variety of subsequent analysis activities. The goal of data wrangling is to make sure that the analysts has access to useful data. The process of data wrangling may include activities such as extracting data in a raw form from several sources, data aggregation, sorting the raw data into useful data structures, and eventually placing the data into a file or files for use in analysis and modeling.

Data cleaning is the process of examining data to detect potential errors, missing data, outliers or unusual values, or other inconsistencies and then correcting the errors or problems that are found. Sometimes errors are the result of recording or transmission problems, and can be corrected by working with the original data source to correct the problem. Effective data cleaning can greatly improve the forecasting process.

Before data are used to develop a time series model, it should be subjected to several different kinds of checks, including but not necessarily limited to the following:

1. Is there missing data?
2. Does the data fall within an expected range?
3. Are there potential outliers or other unusual values?

These types of checks can be automated fairly easily. If this aspect of data cleaning is automated, the rules employed should be periodically evaluated to ensure that they are still appropriate and that changes in the data have not made some of the procedures less effective. However, it is also extremely useful to use graphical displays to assist in identifying unusual data. Techniques such as time series plots, histograms, and scatter diagrams are extremely useful. These and other graphical methods will be described in Chapter 2.

1.4.3 Imputation

Data **imputation** is the process of correcting missing data or replacing outliers with an estimation process. Imputation replaces missing or erroneous values with a "likely" value based on other available information. This enables the analysis to work with statistical techniques which are designed to handle the complete data sets.

Mean value imputation consists of replacing a missing value with the sample average calculated from the nonmissing observations. The big advantage of this method is that it is easy, and if the data does not have any specific trend or seasonal pattern, it leaves the sample mean of the complete data set unchanged. However, one must be careful if there are trends or seasonal patterns, because the sample mean of all of the data may not reflect these patterns. A variation of this is **stochastic mean value imputation**, in which a random variable is added to the mean value to capture some of the noise or variability in the data. The random variable could be assumed to follow a normal distribution with mean zero and standard deviation equal to the standard deviation of the actual observed data. A variation of mean value imputation is to use a subset of the available historical data that reflects any trend or seasonal patterns in the data. For example, consider the time series $y_1, y_2, \ldots, y_T$ and suppose that one observation y_j is missing. We can impute the missing value as

$$y_j^* = \frac{1}{2k}\left(\sum_{t=j-k}^{j-1} y_t + \sum_{t-j+1}^{j+k} y_t\right),$$

where k would be based on the seasonal variability in the data. It is usually chosen as some multiple of the smallest seasonal cycle in the data. So, if the data are monthly and exhibit a monthly cycle, k would be a multiple of 12. **Regression imputation** is a variation of mean value imputation where the imputed value is computed from a model used to predict the missing value. The prediction model does not have to be a linear regression model. For example, it could be a time series model.

Hot deck imputation is an old technique that is also known as the last value carried forward method. The term “hot deck” comes from the use of computer punch cards. The deck of cards was “hot” because it was currently in use. **Cold deck imputation** uses information from a deck of cards not currently in use. In hot deck imputation, the missing values are imputed by using values from similar complete observations. If there are several variables, sort the data by the variables that are most related to the missing observation and then, starting at the top, replace the missing values with the value of the immediately preceding variable. There are many variants of this procedure.

1.5 RESOURCES FOR FORECASTING

There are a variety of good resources that can be helpful to technical professionals involved in developing forecasting models and preparing forecasts. There are three professional journals devoted to forecasting:

- *Journal of Forecasting*
- *International Journal of Forecasting*
- *Journal of Business Forecasting Methods and Systems*

These journals publish a mixture of new methodology, studies devoted to the evaluation of current methods for forecasting, and case studies and applications. In addition to these specialized forecasting journals, there are several other mainstream statistics and operations research/management science journals that publish papers on forecasting, including:

- *Journal of Business and Economic Statistics*
- *Management Science*
- *Naval Research Logistics*
- *Operations Research*
- *International Journal of Production Research*
- *Journal of Applied Statistics*

This is by no means a comprehensive list. Research on forecasting tends to be published in a variety of outlets.

There are several books that are good complements to this one. We recommend Box, Jenkins, and Reinsel (1994); Chatfield (1996); Fuller (1995); Abraham and Ledolter (1983); Montgomery, Johnson, and Gardiner (1990); Wei (2006); and Brockwell and Davis (1991, 2002). Some of these books are more specialized than this one, in that they focus on a specific type of forecasting model such as the autoregressive integrated moving average [ARIMA] model, and some also require more background in statistics and mathematics.

Many statistics software packages have very good capability for fitting a variety of forecasting models. Minitab® Statistical Software, JMP®, the Statistical Analysis System (SAS) and R are the packages that we utilize and illustrate in this book. At the end of most chapters we provide R code for working some of the examples in the chapter. Matlab and S-Plus are also two packages that have excellent capability for solving forecasting problems.

EXERCISES

SS **Student solution available in interactive e-text.**

SS **1.1** Why is forecasting an essential part of the operation of any organization or business?

SS **1.2** What is a time series? Explain the meaning of trend effects, seasonal variations, and random error.

1.3 Explain the difference between a point forecast and an interval forecast.

1.4 What do we mean by a causal forecasting technique?

1.5 Everyone makes forecasts in their daily lives. Identify and discuss a situation where you employ forecasts.

- **a.** What decisions are impacted by your forecasts?
- **b.** How do you evaluate the quality of your forecasts?
- **c.** What is the value to you of a good forecast?
- **d.** What is the harm or penalty associated with a bad forecast?

1.6 What is meant by a rolling horizon forecast?

1.7 Explain the difference between forecast horizon and forecast interval.

1.8 Suppose that you are in charge of capacity planning for a large electric utility. A major part of your job is ensuring that the utility has sufficient generating capacity to meet current and future customer needs. If you do not have enough capacity, you run the risks of brownouts and service interruption. If you have too much capacity, it may cost more to generate electricity.

a. What forecasts do you need to do your job effectively?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

1.9 Your company designs and manufactures apparel for the North American market. Clothing and apparel is a style good, with a relatively limited life. Items not sold at the end of the season are usually sold through off-season outlet and discount retailers. Items not sold through discounting and off-season merchants are often given to charity or sold abroad.

a. What forecasts do you need in this business to be successful?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

d. What are the implications of forecast errors?

1.10 Suppose that you are in charge of production scheduling at a semiconductor manufacturing plant. The plant manufactures about 20 different types of devices, all on 8-inch silicon wafers. Demand for these products varies randomly. When a lot or batch of wafers is started into production, it can take from 4 to 6 weeks before the batch is finished, depending on the type of product. The routing of each batch of wafers through the production tools can be different depending on the type of product.

a. What forecasts do you need in this business to be successful?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

d. Discuss the impact that forecast errors can potentially have on the efficiency with which your factory operates, including work-in-process inventory, meeting customer delivery schedules, and the cycle time to manufacture product.

1.11 You are the administrator of a large metropolitan hospital that operates the only 24-hour emergency room in the area. You must schedule attending physicians, resident physicians, nurses, laboratory, and support personnel to operate this facility effectively.

a. What measures of effectiveness do you think patients use to evaluate the services that you provide?

b. How are forecasts useful to you in planning services that will maximize these measures of effectiveness?

c. What planning horizon do you need to use? Does this lead to short-range or long-range forecasts?

SS **1.12** Consider an airline that operates a network of flights that serves 200 cities in the continental United States. What long-range forecasts do the operators of the airline need to be successful? What forecasting problems does this business face on a daily basis? What are the consequences of forecast errors for the airline?

1.13 Discuss the potential difficulties of forecasting the daily closing price of a specific stock on the New York Stock Exchange. Would the problem be different (harder, easier) if you were asked to forecast the closing price of a group of stocks, all in the same industry (say, the pharmaceutical industry)?

1.14 Explain how large forecast errors can lead to high inventory levels at a retailer; at a manufacturing plant.

1.15 Your company manufactures and distributes soft drink beverages, sold in bottles and cans at retail outlets such as grocery stores, restaurants and other eating/drinking establishments, and vending machines in offices, schools, stores, and other outlets. Your product line includes about 25 different products, and many of these are produced in different package sizes.

a. What forecasts do you need in this business to be successful?

b. Is the demand for your product likely to be seasonal? Explain why or why not?

c. Does the shelf life of your product impact the forecasting problem?

d. What data do you think that you would need to be able to produce successful forecasts?