Wiley Series in Probability and Statistics

Introduction to

TIME SERIES ANALYSIS AND FORECASTING

Second Edition

Douglas C. Montgomery
Cheryl L. Jennings
Murat Kulahci

WILEY

INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING

Second Edition

DOUGLAS C. MONTGOMERY
Arizona State University
Tempe, Arizona, USA

CHERYL L. JENNINGS
Arizona State University
Tempe, Arizona, USA

MURAT KULAHCI
Technical University of Denmark
Lyngby, Denmark
and
Luleå University of Technology
Luleå, Sweden

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data applied for.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE **xi**

1 INTRODUCTION TO FORECASTING **1**

1.1 The Nature and Uses of Forecasts / 1
1.2 Some Examples of Time Series / 6
1.3 The Forecasting Process / 13
1.4 Data for Forecasting / 16
1.4.1 The Data Warehouse / 16
1.4.2 Data Cleaning / 18
1.4.3 Imputation / 18
1.5 Resources for Forecasting / 19
Exercises / 20

2 STATISTICS BACKGROUND FOR FORECASTING **25**

2.1 Introduction / 25
2.2 Graphical Displays / 26
2.2.1 Time Series Plots / 26
2.2.2 Plotting Smoothed Data / 30
2.3 Numerical Description of Time Series Data / 33
2.3.1 Stationary Time Series / 33

2.3.2 Autocovariance and Autocorrelation Functions / 36
2.3.3 The Variogram / 42
2.4 Use of Data Transformations and Adjustments / 46
2.4.1 Transformations / 46
2.4.2 Trend and Seasonal Adjustments / 48
2.5 General Approach to Time Series Modeling and Forecasting / 61
2.6 Evaluating and Monitoring Forecasting Model Performance / 64
2.6.1 Forecasting Model Evaluation / 64
2.6.2 Choosing Between Competing Models / 74
2.6.3 Monitoring a Forecasting Model / 77
2.7 R Commands for Chapter 2 / 84
Exercises / 96

3 REGRESSION ANALYSIS AND FORECASTING 107

3.1 Introduction / 107
3.2 Least Squares Estimation in Linear Regression Models / 110
3.3 Statistical Inference in Linear Regression / 119
3.3.1 Test for Significance of Regression / 120
3.3.2 Tests on Individual Regression Coefficients and Groups of Coefficients / 123
3.3.3 Confidence Intervals on Individual Regression Coefficients / 130
3.3.4 Confidence Intervals on the Mean Response / 131
3.4 Prediction of New Observations / 134
3.5 Model Adequacy Checking / 136
3.5.1 Residual Plots / 136
3.5.2 Scaled Residuals and PRESS / 139
3.5.3 Measures of Leverage and Influence / 144
3.6 Variable Selection Methods in Regression / 146
3.7 Generalized and Weighted Least Squares / 152
3.7.1 Generalized Least Squares / 153
3.7.2 Weighted Least Squares / 156
3.7.3 Discounted Least Squares / 161
3.8 Regression Models for General Time Series Data / 177

3.8.1 Detecting Autocorrelation: The Durbin–Watson Test / 178
3.8.2 Estimating the Parameters in Time Series Regression Models / 184
3.9 Econometric Models / 205
3.10 R Commands for Chapter 3 / 209
Exercises / 219

4 EXPONENTIAL SMOOTHING METHODS 233

4.1 Introduction / 233
4.2 First-Order Exponential Smoothing / 239
4.2.1 The Initial Value, $\tilde{y}_0$ / 241
4.2.2 The Value of λ / 241
4.3 Modeling Time Series Data / 245
4.4 Second-Order Exponential Smoothing / 247
4.5 Higher-Order Exponential Smoothing / 257
4.6 Forecasting / 259
4.6.1 Constant Process / 259
4.6.2 Linear Trend Process / 264
4.6.3 Estimation of σ_e^2 / 273
4.6.4 Adaptive Updating of the Discount Factor / 274
4.6.5 Model Assessment / 276
4.7 Exponential Smoothing for Seasonal Data / 277
4.7.1 Additive Seasonal Model / 277
4.7.2 Multiplicative Seasonal Model / 280
4.8 Exponential Smoothing of Biosurveillance Data / 286
4.9 Exponential Smoothers and Arima Models / 299
4.10 R Commands for Chapter 4 / 300
Exercises / 311

5 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS 327

5.1 Introduction / 327
5.2 Linear Models for Stationary Time Series / 328
5.2.1 Stationarity / 329
5.2.2 Stationary Time Series / 329

5.3 Finite Order Moving Average Processes / 333
5.3.1 The First-Order Moving Average Process, MA(1) / 334
5.3.2 The Second-Order Moving Average Process, MA(2) / 336
5.4 Finite Order Autoregressive Processes / 337
5.4.1 First-Order Autoregressive Process, AR(1) / 338
5.4.2 Second-Order Autoregressive Process, AR(2) / 341
5.4.3 General Autoregressive Process, AR(*p*) / 346
5.4.4 Partial Autocorrelation Function, PACF / 348
5.5 Mixed Autoregressive–Moving Average Processes / 354
5.5.1 Stationarity of ARMA(*p, q*) Process / 355
5.5.2 Invertibility of ARMA(*p, q*) Process / 355
5.5.3 ACF and PACF of ARMA(*p, q*) Process / 356
5.6 Nonstationary Processes / 363
5.6.1 Some Examples of ARIMA(*p, d, q*) Processes / 363
5.7 Time Series Model Building / 367
5.7.1 Model Identification / 367
5.7.2 Parameter Estimation / 368
5.7.3 Diagnostic Checking / 368
5.7.4 Examples of Building ARIMA Models / 369
5.8 Forecasting Arima Processes / 378
5.9 Seasonal Processes / 383
5.10 Arima Modeling of Biosurveillance Data / 393
5.11 Final Comments / 399
5.12 R Commands for Chapter 5 / 401
Exercises / 412

6 TRANSFER FUNCTIONS AND INTERVENTION MODELS 427

6.1 Introduction / 427
6.2 Transfer Function Models / 428
6.3 Transfer Function–Noise Models / 436
6.4 Cross-Correlation Function / 436
6.5 Model Specification / 438
6.6 Forecasting with Transfer Function–Noise Models / 456

6.7 Intervention Analysis / 462
6.8 R Commands for Chapter 6 / 473
Exercises / 486

7 SURVEY OF OTHER FORECASTING METHODS 493

7.1 Multivariate Time Series Models and Forecasting / 493
7.1.1 Multivariate Stationary Process / 494
7.1.2 Vector ARIMA Models / 494
7.1.3 Vector AR (VAR) Models / 496
7.2 State Space Models / 502
7.3 Arch and Garch Models / 507
7.4 Direct Forecasting of Percentiles / 512
7.5 Combining Forecasts to Improve Prediction Performance / 518
7.6 Aggregation and Disaggregation of Forecasts / 522
7.7 Neural Networks and Forecasting / 526
7.8 Spectral Analysis / 529
7.9 Bayesian Methods in Forecasting / 535
7.10 Some Comments on Practical Implementation and Use of Statistical Forecasting Procedures / 542
7.11 R Commands for Chapter 7 / 545
Exercises / 550

APPENDIX A STATISTICAL TABLES 561

APPENDIX B DATA SETS FOR EXERCISES 581

APPENDIX C INTRODUCTION TO R 627

BIBLIOGRAPHY 631

INDEX 639

PREFACE

Analyzing time-oriented data and forecasting future values of a time series are among the most important problems that analysts face in many fields, ranging from finance and economics to managing production operations, to the analysis of political and social policy sessions, to investigating the impact of humans and the policy decisions that they make on the environment. Consequently, there is a large group of people in a variety of fields, including finance, economics, science, engineering, statistics, and public policy who need to understand some basic concepts of time series analysis and forecasting. Unfortunately, most basic statistics and operations management books give little if any attention to time-oriented data and little guidance on forecasting. There are some very good high level books on time series analysis. These books are mostly written for technical specialists who are taking a doctoral-level course or doing research in the field. They tend to be very theoretical and often focus on a few specific topics or techniques. We have written this book to fill the gap between these two extremes.

We have made a number of changes in this revision of the book. New material has been added on data preparation for forecasting, including dealing with outliers and missing values, use of the variogram and sections on the spectrum, and an introduction to Bayesian methods in forecasting. We have added many new exercises and examples, including new data sets in Appendix B, and edited many sections of the text to improve the clarity of the presentation.

Like the first edition, this book is intended for practitioners who make real-world forecasts. We have attempted to keep the mathematical level modest to encourage a variety of users for the book. Our focus is on short- to medium-term forecasting where statistical methods are useful. Since many organizations can improve their effectiveness and business results by making better short- to medium-term forecasts, this book should be useful to a wide variety of professionals. The book can also be used as a textbook for an applied forecasting and time series analysis course at the advanced undergraduate or first-year graduate level. Students in this course could come from engineering, business, statistics, operations research, mathematics, computer science, and any area of application where making forecasts is important. Readers need a background in basic statistics (previous exposure to linear regression would be helpful, but not essential), and some knowledge of matrix algebra, although matrices appear mostly in the chapter on regression, and if one is interested mainly in the results, the details involving matrix manipulation can be skipped. Integrals and derivatives appear in a few places in the book, but no detailed working knowledge of calculus is required.

Successful time series analysis and forecasting requires that the analyst interact with computer software. The techniques and algorithms are just not suitable to manual calculations. We have chosen to demonstrate the techniques presented using three packages: Minitab®, JMP®, and R, and occasionally SAS®. We have selected these packages because they are widely used in practice and because they have generally good capability for analyzing time series data and generating forecasts. Because R is increasingly popular in statistics courses, we have included a section in each chapter showing the R code necessary for working some of the examples in the chapter. We have also added a brief appendix on the use of R. The basic principles that underlie most of our presentation are not specific to any particular software package. Readers can use any software that they like or have available that has basic statistical forecasting capability. While the text examples do utilize these particular software packages and illustrate some of their features and capability, these features or similar ones are found in many other software packages.

There are three basic approaches to generating forecasts: regression-based methods, heuristic smoothing methods, and general time series models. Because all three of these basic approaches are useful, we give an introduction to all of them. Chapter 1 introduces the basic forecasting problem, defines terminology, and illustrates many of the common features of time series data. Chapter 2 contains many of the basic statistical tools used in analyzing time series data. Topics include plots, numerical summaries of time series data including the autocovariance and autocorrelation functions, transformations, differencing, and decomposing a time series into trend and seasonal components. We also introduce metrics for evaluating forecast errors and methods for evaluating and tracking forecasting performance over time. Chapter 3 discusses regression analysis and its use in forecasting. We discuss both crosssection and time series regression data, least squares and maximum likelihood model fitting, model adequacy checking, prediction intervals, and weighted and generalized least squares. The first part of this chapter covers many of the topics typically seen in an introductory treatment of regression, either in a stand-alone course or as part of another applied statistics course. It should be a reasonable review for many readers. Chapter 4 presents exponential smoothing techniques, both for time series with polynomial components and for seasonal data. We discuss and illustrate methods for selecting the smoothing constant(s), forecasting, and constructing prediction intervals. The explicit time series modeling approach to forecasting that we have chosen to emphasize is the autoregressive integrated moving average (ARIMA) model approach. Chapter 5 introduces ARIMA models and illustrates how to identify and fit these models for both nonseasonal and seasonal time series. Forecasting and prediction interval construction are also discussed and illustrated. Chapter 6 extends this discussion into transfer function models and intervention modeling and analysis. Chapter 7 surveys several other useful topics from time series analysis and forecasting, including multivariate time series problems, ARCH and GARCH models, and combinations of forecasts. We also give some practical advice for using statistical approaches to forecasting and provide some information about realistic expectations. The last two chapters of the book are somewhat higher in level than the first five.

Each chapter has a set of exercises. Some of these exercises involve analyzing the data sets given in Appendix B. These data sets represent an interesting cross section of real time series data, typical of those encountered in practical forecasting problems. Most of these data sets are used in exercises in two or more chapters, an indication that there are usually several approaches to analyzing, modeling, and forecasting a time series. There are other good sources of data for practicing the techniques given in this book. Some of the ones that we have found very interesting and useful include the U.S. Department of Labor—Bureau of Labor Statistics (http://www.bls.gov/data/home.htm), the U.S. Department of Agriculture—National Agricultural Statistics Service, Quick Stats Agricultural Statistics Data (http://www.nass.usda.gov/Data_and_Statistics/Quick_Stats/index.asp), the U.S. Census Bureau (http://www.census.gov), and the U.S.

Department of the Treasury (http://www.treas.gov/offices/domestic-finance/debt-management/interest-rate/). The time series data library created by Rob Hyndman at Monash University (http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/index.htm) and the time series data library at the Mathematics Department of the University of York (http://www.york.ac.uk/depts/maths/data/ts/) also contain many excellent data sets. Some of these sources provide links to other data. Data sets and other materials related to this book can be found at ftp://ftp.wiley.com/public/scitechmed/ timeseries.

We would like to thank the many individuals who provided feedback and suggestions for improvement to the first edition. We found these suggestions most helpful. We are indebted to Clifford Long who generously provided the R codes he used with his students when he taught from the book. We found his codes very helpful in putting the end-of-chapter R code sections together. We also have placed a premium in the book on bridging the gap between theory and practice. We have not emphasized proofs or technical details and have tried to give intuitive explanations of the material whenever possible. The result is a book that can be used with a wide variety of audiences, with different interests and technical backgrounds, whose common interests are understanding how to analyze time-oriented data and constructing good short-term statistically based forecasts.

We express our appreciation to the individuals and organizations who have given their permission to use copyrighted material. These materials are noted in the text.

Douglas C. Montgomery
Cheryl L. Jennings
Murat Kulahci

CHAPTER 1

INTRODUCTION TO 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.

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

Forecasting is an important problem that spans many fields including business and industry, government, economics, environmental sciences, 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 December 2006 (data in Appendix B, Table B.1). This graph is called a **time**

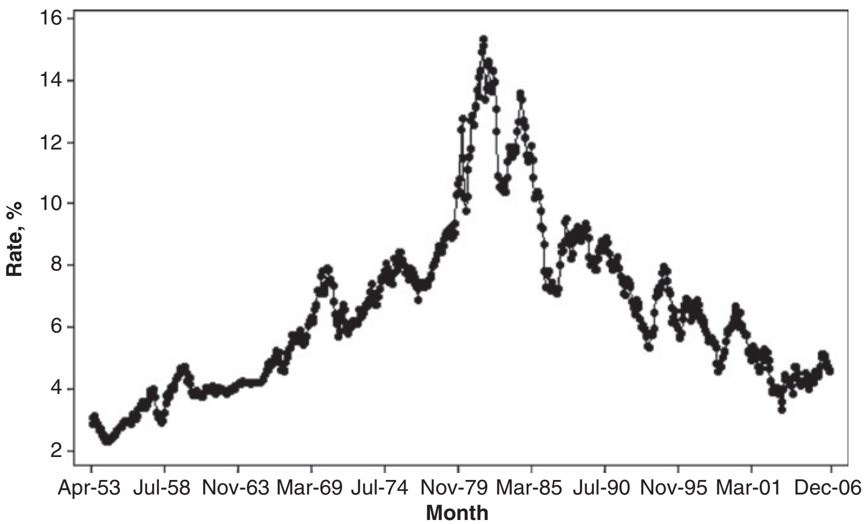

FIGURE 1.1 Time series plot of the market yield on US Treasury Securities at 10-year constant maturity. *Source:* US Treasury.

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.

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

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 1.2 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.

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

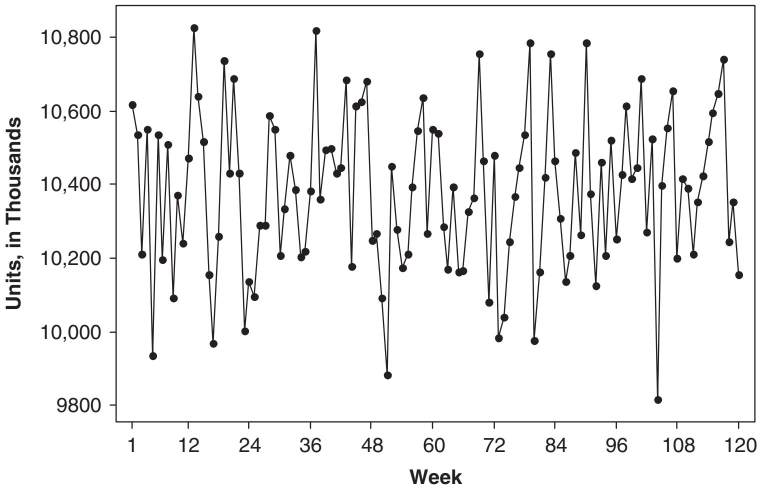

FIGURE 1.2 Pharmaceutical product sales.

average, while a value below the average tends to be followed by other values below the average.

The viscosity readings plotted in Figure 1.3 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.

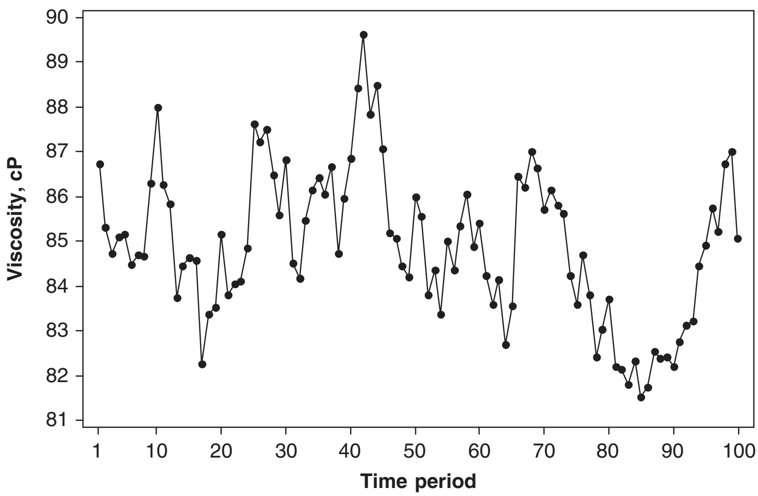

FIGURE 1.3 Chemical process viscosity readings.

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 1.4 (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.

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 1.5) 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

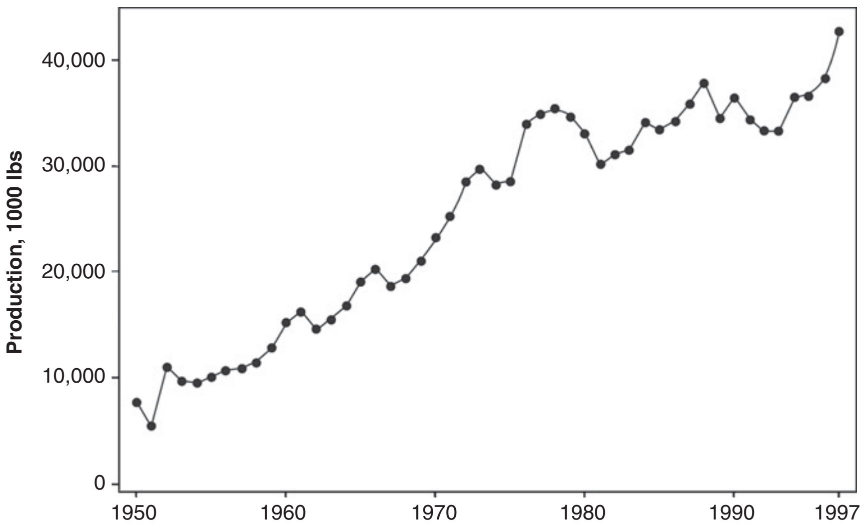

FIGURE 1.4 The US annual production of blue and gorgonzola cheeses. *Source:* USDA–NASS.

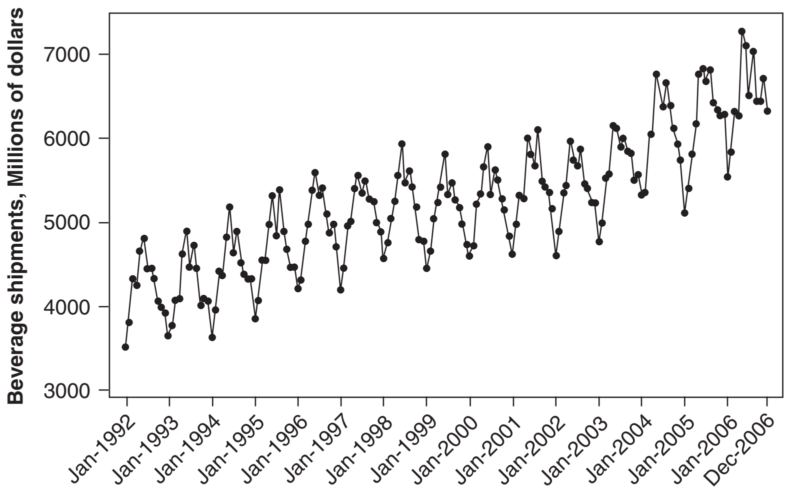

FIGURE 1.5 The US beverage manufacturer monthly product shipments, unadjusted. *Source:* US Census Bureau.

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."

The plot of the annual mean anomaly in global surface air temperature (Figure 1.6) 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 1.7) 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

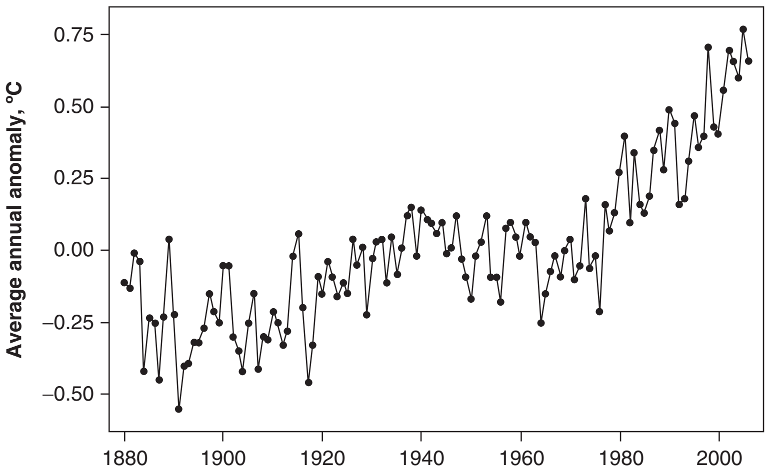

FIGURE 1.6 Global mean surface air temperature annual anomaly. *Source:* NASA-GISS.

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

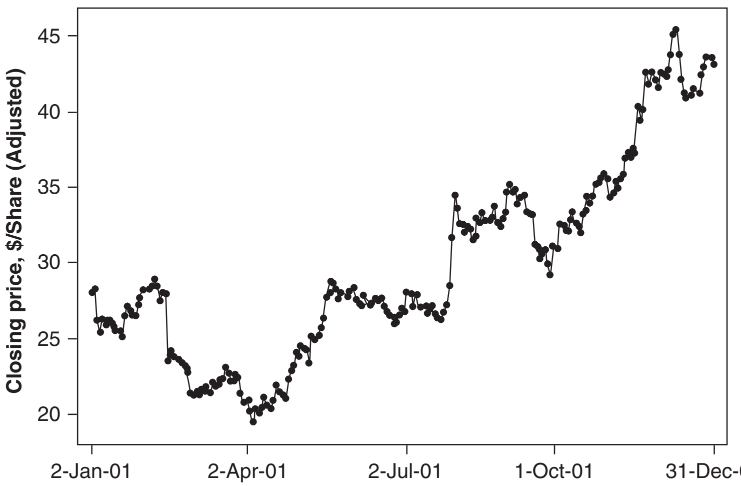

FIGURE 1.7 Whole foods market stock price, daily closing adjusted for splits.

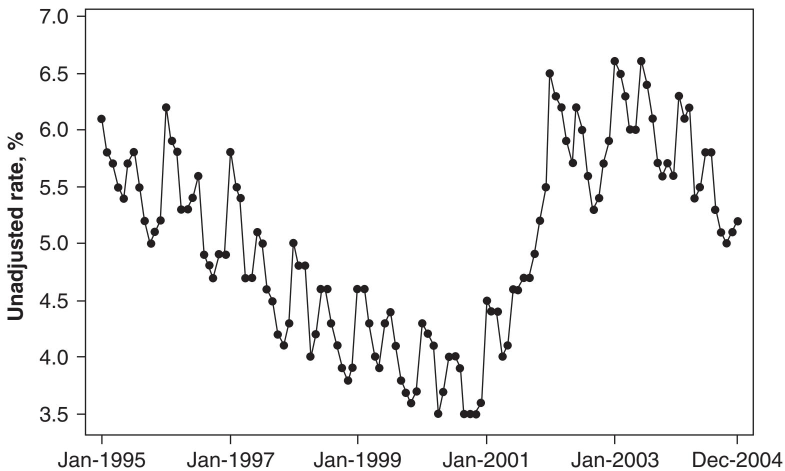

FIGURE 1.8 Monthly unemployment rate—full-time labor force, unadjusted. *Source:* US Department of Labor-BLS.

on the employment situation, for projections with impact on hiring and training, and for a multitude of other business planning activities. The data 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 1.8) exhibits a mixture of patterns, similar to Figure 1.5 (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 1.9, 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

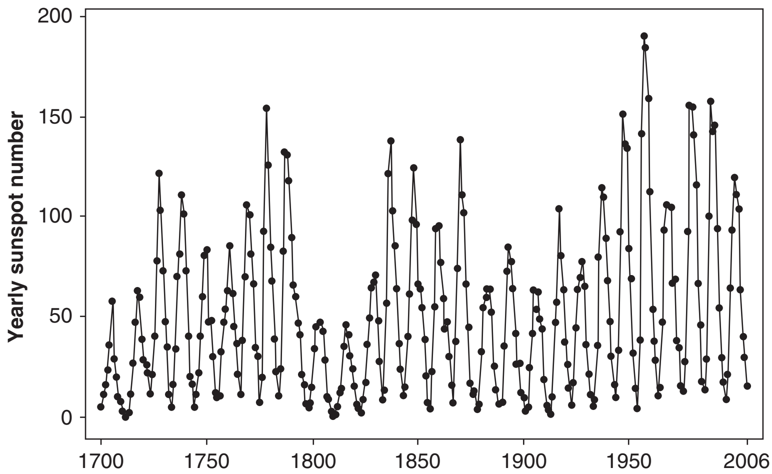

FIGURE 1.9 The international sunspot number. *Source:* SIDC.

availability resulting from a fire at one of the four production facilities. The 5-week reduction is apparent in the time series plot of weekly sales shown in Figure 1.10.

Another type of unusual event may be the failure of the data measurement or collection system. After recording a vastly different viscosity reading at time period 70 (Figure 1.11), the measurement system was

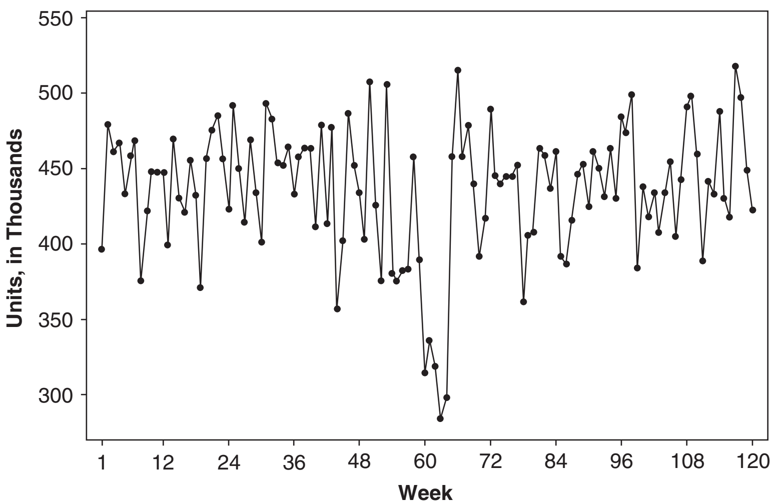

FIGURE 1.10 Pharmaceutical product sales.

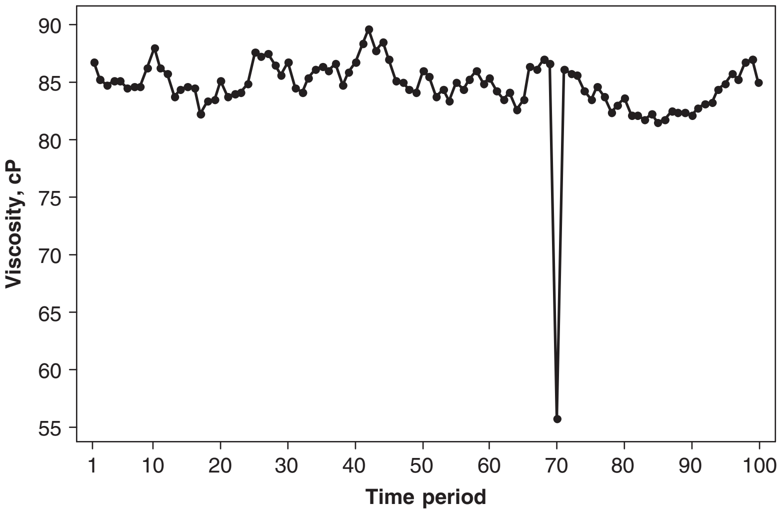

FIGURE 1.11 Chemical process viscosity readings, with sensor malfunction.

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

These activities are shown in Figure 1.12.

Problem definition involves developing understanding of how the forecast will be used along with the expectations of the "customer" (the user of

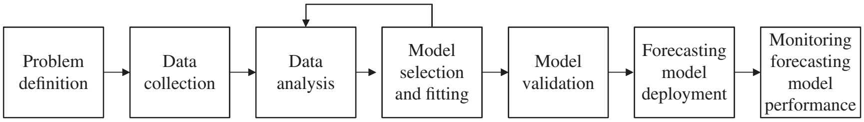

FIGURE 1.12 The forecasting process.

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**

cleaning. We will discuss some of the important data cleaning 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 Cleaning

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

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

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?

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?

CHAPTER 2

STATISTICS BACKGROUND FOR FORECASTING

The future ain't what it used to be.

YOGI BERRA, *New York Yankees catcher*

2.1 INTRODUCTION

This chapter presents some basic statistical methods essential to modeling, analyzing, and forecasting time series data. Both graphical displays and numerical summaries of the properties of time series data are presented. We also discuss the use of data transformations and adjustments in forecasting and some widely used methods for characterizing and monitoring the performance of a forecasting model. Some aspects of how these performance measures can be used to select between competing forecasting techniques are also presented.

Forecasts are based on data or observations on the variable of interest. These data are usually in the form of a **time series**. Suppose that there are T periods of data available, with period T being the most recent. We will let the observation on this variable at time period t be denoted by y_t, $t = 1, 2, \ldots, T$. This variable can represent a cumulative quantity, such as the

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

total demand for a product during period t, or an instantaneous quantity, such as the daily closing price of a specific stock on the New York Stock Exchange.

Generally, we will need to distinguish between a **forecast** or **predicted value** of y_t that was made at some previous time period, say, $t - \tau$, and a **fitted value** of y_t that has resulted from estimating the parameters in a time series model to historical data. Note that τ is the forecast lead time. The forecast made at time period $t - \tau$ is denoted by $\hat{y}_t(t - \tau)$. There is a lot of interest in the **lead − 1** forecast, which is the forecast of the observation in period t, y_t, made one period prior, $\hat{y}_t(t - 1)$. We will denote the fitted value of y_t by $\hat{y}_t$.

We will also be interested in analyzing **forecast errors**. The forecast error that results from a forecast of y_t that was made at time period $t - \tau$ is the **lead − τ forecast error**

$$e_t(\tau) = y_t - \hat{y}_t(t - \tau). \tag{2.1}$$

For example, the lead − 1 forecast error is

$$e_t(1) = y_t - \hat{y}_t(t - 1).$$

The difference between the observation y_t and the value obtained by fitting a time series model to the data, or a fitted value $\hat{y}_t$ defined earlier, is called a **residual**, and is denoted by

$$e_t = y_t - \hat{y}_t. \tag{2.2}$$

The reason for this careful distinction between forecast errors and residuals is that models usually fit historical data better than they forecast. That is, the residuals from a model-fitting process will almost always be smaller than the forecast errors that are experienced when that model is used to forecast future observations.

2.2 GRAPHICAL DISPLAYS

2.2.1 Time Series Plots

Developing a forecasting model should always begin with graphical display and analysis of the available data. Many of the broad general features of a time series can be seen visually. This is not to say that analytical tools are not useful, because they are, but the human eye can be a very sophisticated data analysis tool. To paraphrase the great New York Yankees catcher Yogi Berra, "You can observe a lot just by watching."

The basic graphical display for time series data is the **time series plot**, illustrated in Chapter 1. This is just a graph of y_t versus the time period, t, for $t = 1, 2, \ldots, T$. Features such as trend and seasonality are usually easy to see from the time series plot. It is interesting to observe that some of the classical tools of descriptive statistics, such as the histogram and the stem-and-leaf display, are not particularly useful for time series data because they do not take time order into account.

Example 2.1 Figures 2.1 and 2.2 show time series plots for viscosity readings and beverage production shipments (originally shown in Figures 1.3 and 1.5, respectively). At the right-hand side of each time series plot is a histogram of the data. Note that while the two time series display very different characteristics, the histograms are remarkably similar. Essentially, the histogram summarizes the data across the time dimension, and in so doing, the key time-dependent features of the data are lost. Stem-and-leaf plots and boxplots would have the same issues, losing time-dependent features.

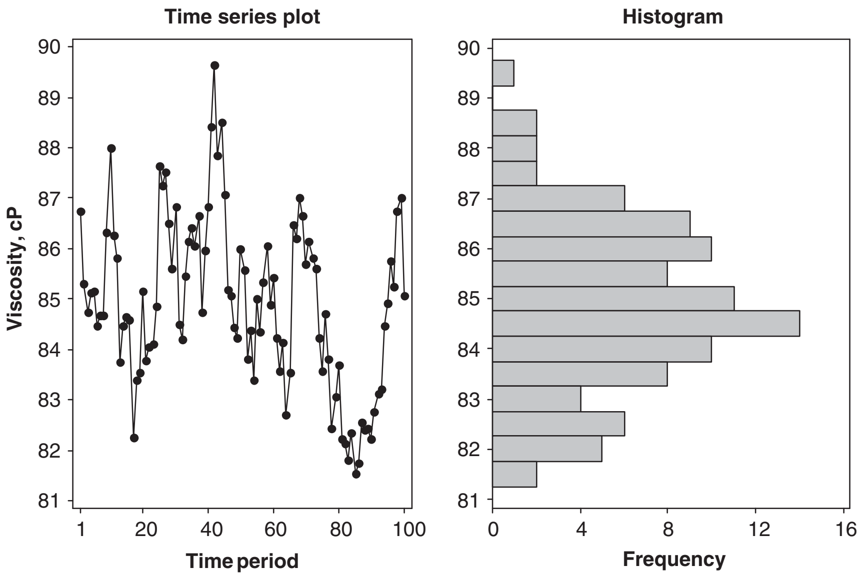

FIGURE 2.1 Time series plot and histogram of chemical process viscosity readings.

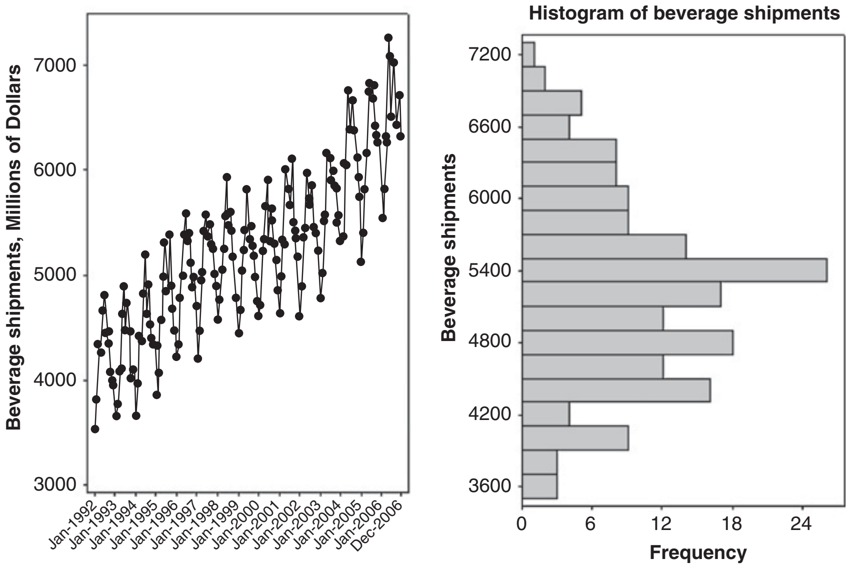

FIGURE 2.2 Time series plot and histogram of beverage production shipments.

When there are two or more variables of interest, **scatter plots** can be useful in displaying the relationship between the variables. For example, Figure 2.3 is a scatter plot of the annual global mean surface air temperature anomaly first shown in Figure 1.6 versus atmospheric CO_2 concentrations. The scatter plot clearly reveals a relationship between the two variables:

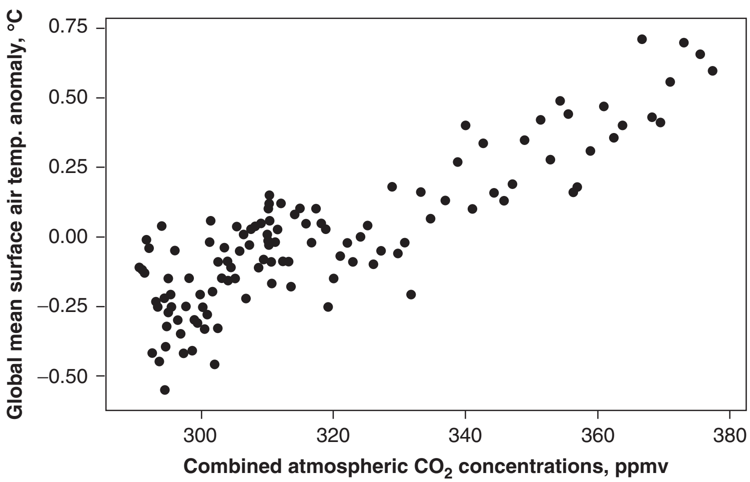

FIGURE 2.3 Scatter plot of temperature anomaly versus CO_2 concentrations. *Sources*: NASA–GISS (anomaly), DOE–DIAC (CO_2).

low concentrations of CO_2 are usually accompanied by negative anomalies, and higher concentrations of CO_2 tend to be accompanied by positive anomalies. Note that this does not imply that higher concentrations of CO_2 actually *cause* higher temperatures. The scatter plot cannot establish a causal relationship between two variables (neither can naive statistical modeling techniques, such as regression), but it is useful in displaying how the variables have varied together in the historical data set.

There are many variations of the time series plot and other graphical displays that can be constructed to show specific features of a time series. For example, Figure 2.4 displays daily price information for Whole Foods Market stock during the first quarter of 2001 (the trading days from January 2, 2001 through March 30, 2001). This chart, created in Excel®, shows the opening, closing, highest, and lowest prices experienced within a trading day for the first quarter. If the opening price was higher than the closing price, the box is filled, whereas if the closing price was higher than the opening price, the box is open. This type of plot is potentially more useful than a time series plot of just the closing (or opening) prices, because it shows the volatility of the stock within a trading day. The volatility of an asset is often of interest to investors because it is a measure of the inherent risk associated with the asset.

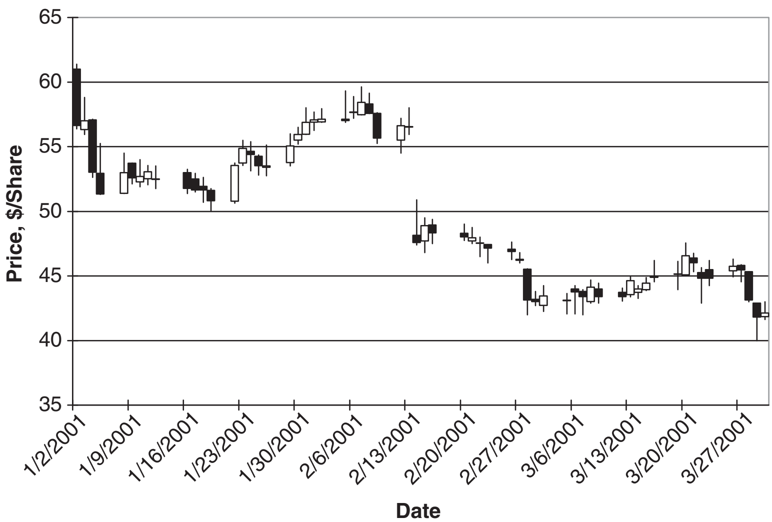

FIGURE 2.4 Open-high/close-low chart of Whole Foods Market stock price. *Source*: `finance.yahoo.com`.

2.2.2 Plotting Smoothed Data

Sometimes it is useful to overlay a **smoothed** version of the original data on the original time series plot to help reveal patterns in the original data. There are several types of data smoothers that can be employed. One of the simplest and most widely used is the ordinary or simple moving average. A simple **moving average** of span N assigns weights $1/N$ to the most recent N observations $y_T, y_{T-1}, \dots, y_{T-N+1}$, and weight zero to all other observations. If we let M_T be the moving average, then the N-span moving average at time period T is

$$M_T = \frac{y_T + y_{T-1} + \cdots + y_{T-N+1}}{N} = \frac{1}{N} \sum_{t=T-N+1}^{T} y_t \tag{2.3}$$

Clearly, as each new observation becomes available it is added into the sum from which the moving average is computed and the oldest observation is discarded. The moving average has less variability than the original observations; in fact, if the variance of an individual observation y_t is σ^2, then assuming that the observations are uncorrelated the variance of the moving average is

$$\mathrm{Var}(M_T) = \mathrm{Var}\left(\frac{1}{N} \sum_{t=T-N+1}^{N} y_t\right) = \frac{1}{N^2} \sum_{t=T-N+1}^{N} \mathrm{Var}(y_t) = \frac{\sigma^2}{N}$$

Sometimes a "centered" version of the moving average is used, such as in

$$M_t = \frac{1}{S+1} \sum_{i=-S}^{S} y_{t-i} \tag{2.4}$$

where the span of the centered moving average is $N = 2S + 1$.

Example 2.2 Figure 2.5 plots the annual global mean surface air temperature anomaly data along with a five-period (a period is 1 year) moving average of the same data. Note that the moving average exhibits less variability than found in the original series. It also makes some features of the data easier to see; for example, it is now more obvious that the global air temperature decreased from about 1940 until about 1975.

Plots of moving averages are also used by analysts to evaluate stock price trends; common MA periods are 5, 10, 20, 50, 100, and 200 days. A time series plot of Whole Foods Market stock price with a 50-day moving

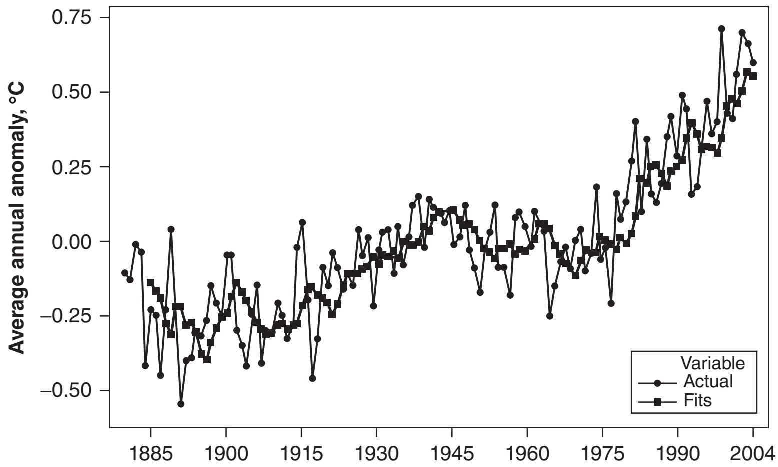

FIGURE 2.5 Time series plot of global mean surface air temperature anomaly, with five-period moving average. *Source*: NASA–GISS.

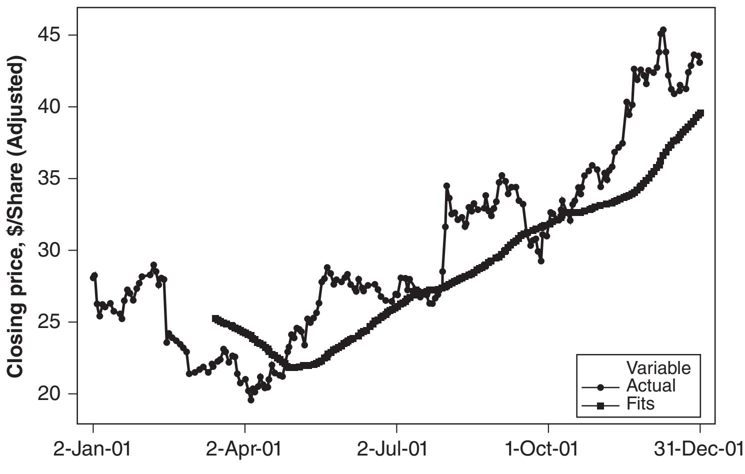

FIGURE 2.6 Time series plot of Whole Foods Market stock price, with 50-day moving average. *Source*: `finance.yahoo.com`.

average is shown in Figure 2.6. The moving average plot smoothes the day-to-day noise and shows a generally increasing trend.

The simple moving average is a **linear data smoother**, or a **linear filter**, because it replaces each observation y_t with a linear combination of the other data points that are near to it in time. The weights in the linear combination are equal, so the linear combination here is an average. Of course, unequal weights could be used. For example, the **Hanning filter** is a weighted, centered moving average

$$M_t^{\mathrm{H}} = 0.25y_{t+1} + 0.5y_t + 0.25y_{t-1}$$

Julius von Hann, a nineteenth century Austrian meteorologist, used this filter to smooth weather data.

An obvious disadvantage of a linear filter such as a moving average is that an unusual or erroneous data point or an outlier will dominate the moving averages that contain that observation, contaminating the moving averages for a length of time equal to the span of the filter. For example, consider the sequence of observations

$$15, 18, 13, 12, 16, 14, 16, 17, 18, 15, 18, 200, 19, 14, 21, 24, 19, 25$$

which increases reasonably steadily from 15 to 25, except for the unusual value 200. Any reasonable smoothed version of the data should also increase steadily from 15 to 25 and not emphasize the value 200. Now even if the value 200 is a legitimate observation, and not the result of a data recording or reporting error (perhaps it should be 20!), it is so unusual that it deserves special attention and should likely not be analyzed along with the rest of the data.

Odd-span **moving medians** (also called **running medians**) are an alternative to moving averages that are effective data smoothers when the time series may be contaminated with unusual values or outliers. The moving median of span N is defined as

$$m_t^{[N]} = med(y_{t-u}, \ldots, y_t, \ldots, y_{t+u}), \tag{2.5}$$

where $N = 2u + 1$. The median is the middle observation in rank order (or order of value). The moving median of span 3 is a very popular and effective data smoother, where

$$m_t^{[3]} = med(y_{t-1}, y_t, y_{t+1}).$$

This smoother would process the data three values at a time, and replace the three original observations by their median. If we apply this smoother to the data above, we obtain

$$\_\_\_, 15, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, 24, \_\_\_.$$

This smoothed data are a reasonable representation of the original data, but they conveniently ignore the value 200. The end values are lost when using the moving median, and they are represented by "____".

In general, a moving median will pass monotone sequences of data unchanged. It will follow a step function in the data, but it will eliminate a spike or more persistent upset in the data that has duration of at most u consecutive observations. Moving medians can be applied more than once if desired to obtain an even smoother series of observations. For example, applying the moving median of span 3 to the smoothed data above results in

$$\_\_\_\_, \_\_\_\_, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, \_\_\_\_, \_\_\_\_.$$

These data are now as smooth as it can get; that is, repeated application of the moving median will not change the data, apart from the end values.

If there are a lot of observations, the information loss from the missing end values is not serious. However, if it is necessary or desirable to keep the lengths of the original and smoothed data sets the same, a simple way to do this is to "copy on" or add back the end values from the original data. This would result in the smoothed data:

$$15, 18, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, 19, 25$$

There are also methods for smoothing the end values. Tukey (1979) is a basic reference on this subject and contains many other clever and useful techniques for data analysis.

Example 2.3 The chemical process viscosity readings shown in Figure 1.11 are an example of a time series that benefits from smoothing to evaluate patterns. The selection of a moving median over a moving average, as shown in Figure 2.7, minimizes the impact of the invalid measurements, such as the one at time period 70.

2.3 NUMERICAL DESCRIPTION OF TIME SERIES DATA

2.3.1 Stationary Time Series

A very important type of time series is a **stationary** time series. A time series is said to be **strictly stationary** if its properties are not affected

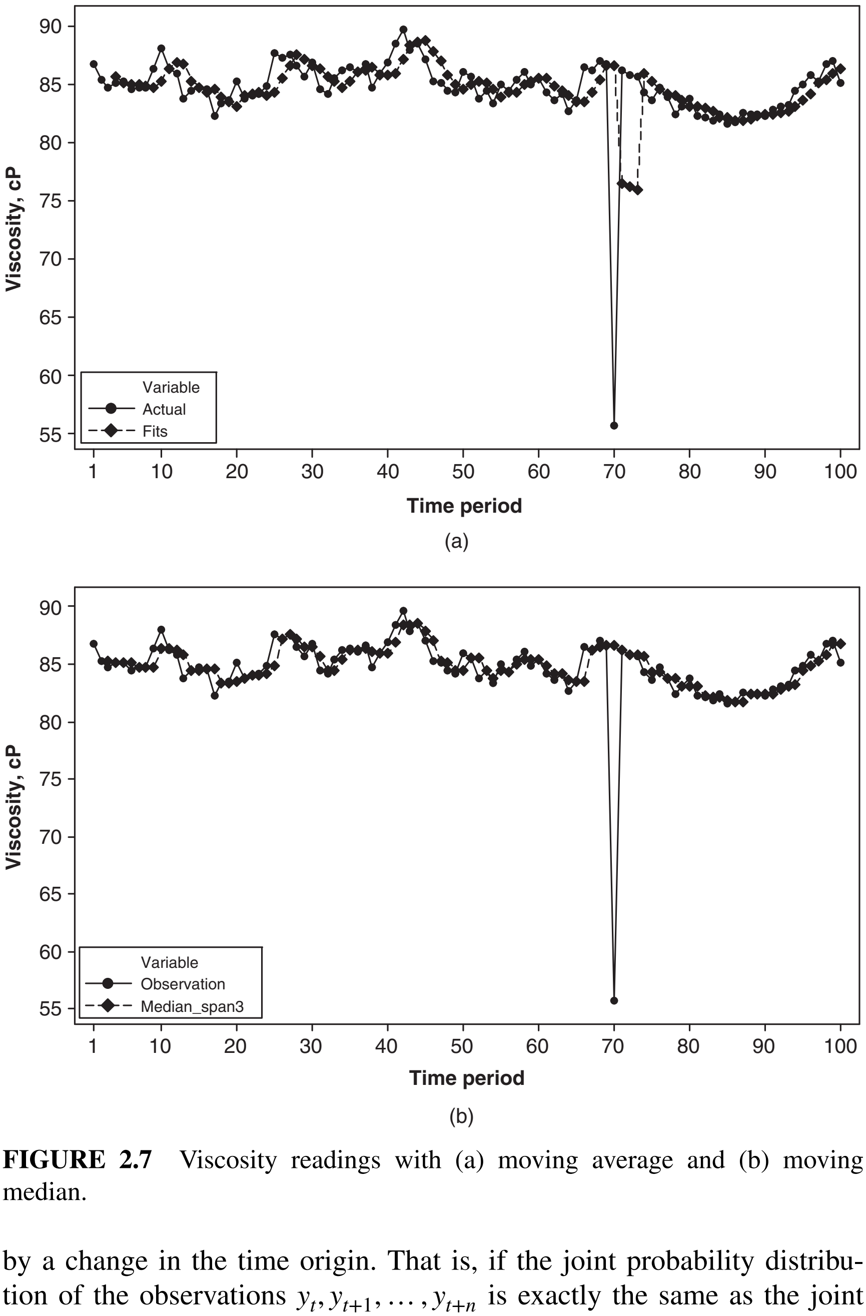

FIGURE 2.7 Viscosity readings with (a) moving average and (b) moving median.

by a change in the time origin. That is, if the joint probability distribution of the observations $y_t, y_{t+1}, \dots, y_{t+n}$ is exactly the same as the joint probability distribution of the observations $y_{t+k}, y_{t+k+1}, \dots, y_{t+k+n}$ then the time series is strictly stationary. When $n = 0$ the stationarity assumption means that the probability distribution of y_t is the same for all time periods

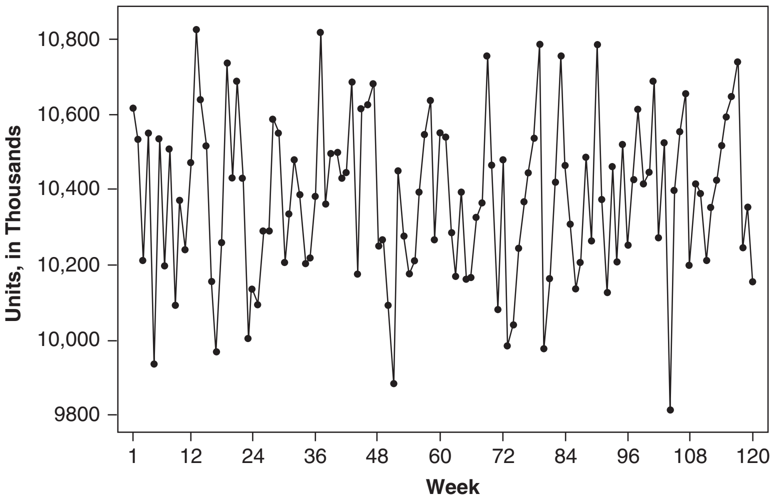

FIGURE 2.8 Pharmaceutical product sales.

and can be written as $f(y)$. The pharmaceutical product sales and chemical viscosity readings time series data originally shown in Figures 1.2 and 1.3, respectively, are examples of stationary time series. The time series plots are repeated in Figures 2.8 and 2.9 for convenience. Note that both time series seem to vary around a fixed level. Based on the earlier definition, this is a characteristic of stationary time series. On the other hand, the Whole

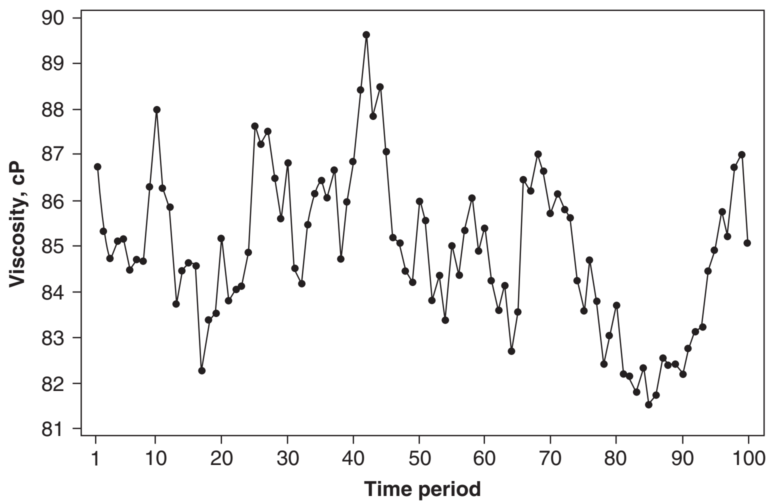

FIGURE 2.9 Chemical process viscosity readings.

Foods Market stock price data in Figure 1.7 tends to wander around or drift, with no obvious fixed level. This is behavior typical of a nonstationary time series.

Stationary implies a type of statistical **equilibrium** or **stability** in the data. Consequently, the time series has a constant mean defined in the usual way as

$$\mu_y = E(y) = \int_{-\infty}^{\infty} yf(y)dy \tag{2.6}$$

and constant variance defined as

$$\sigma_y^2 = \text{Var}(y) = \int_{-\infty}^{\infty} (y - \mu_y)^2 f(y)dy. \tag{2.7}$$

The sample mean and sample variance are used to estimate these parameters. If the observations in the time series are $y_1, y_2, \ldots, y_T$, then the sample mean is

$$\bar{y} = \hat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{2.8}$$

and the sample variance is

$$s^2 = \hat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2. \tag{2.9}$$

Note that the divisor in Eq. (2.9) is T rather than the more familiar $T - 1$. This is the common convention in many time series applications, and because T is usually not small, there will be little difference between using T instead of $T - 1$.

2.3.2 Autocovariance and Autocorrelation Functions

If a time series is stationary this means that the joint probability distribution of any two observations, say, y_t and y_{t+k}, is the same for any two time periods t and $t + k$ that are separated by the same interval k. Useful information about this joint distribution, and hence about the nature of the time series, can be obtained by plotting a scatter diagram of all of the data pairs y_t, y_{t+k} that are separated by the same interval k. The interval k is called the **lag**.

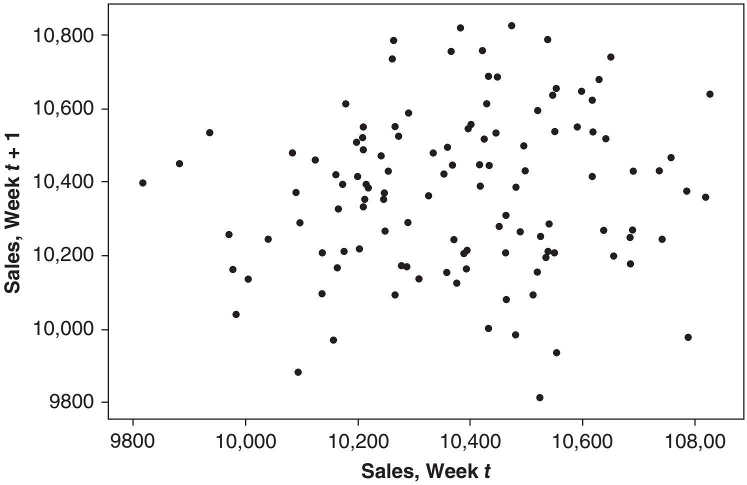

FIGURE 2.10 Scatter diagram of pharmaceutical product sales at lag $k = 1$.

Example 2.4 Figure 2.10 is a scatter diagram for the pharmaceutical product sales for lag $k = 1$ and Figure 2.11 is a scatter diagram for the chemical viscosity readings for lag $k = 1$. Both scatter diagrams were constructed by plotting y_{t+1} versus y_t. Figure 2.10 exhibits little structure; the plotted pairs of adjacent observations y_t, y_{t+1} seem to be **uncorrelated**. That is, the value of y in the current period does not provide any useful information about the value of y that will be observed in the next period. A different story is revealed in Figure 2.11, where we observe that the

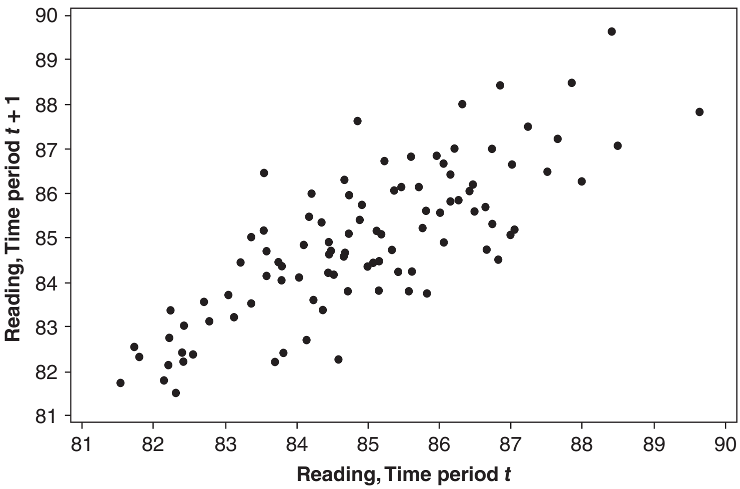

FIGURE 2.11 Scatter diagram of chemical viscosity readings at lag $k = 1$.

pairs of adjacent observations y_{t+1}, y_t are **positively correlated**. That is, a small value of y tends to be followed in the next time period by another small value of y, and a large value of y tends to be followed immediately by another large value of y. Note from inspection of Figures 2.10 and 2.11 that the behavior inferred from inspection of the scatter diagrams is reflected in the observed time series.

The covariance between y_t and its value at another time period, say, y_{t+k} is called the **autocovariance** at lag k, defined by

$$\gamma_k = \text{Cov}(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)]. \tag{2.10}$$

The collection of the values of $\gamma_k, k = 0, 1, 2, \ldots$ is called the **autocovariance function**. Note that the autocovariance at lag $k = 0$ is just the variance of the time series; that is, $\gamma_0 = \sigma_y^2$,which is constant for a stationary time series. The **autocorrelation coefficient** at lag k for a stationary time series is

$$\rho_k = \frac{E[(y_t - \mu)(y_{t+k} - \mu)]}{\sqrt{E[(y_t - \mu)^2]E[(y_{t+k} - \mu)^2]}} = \frac{\text{Cov}(y_t, y_{t+k})}{\text{Var}(y_t)} = \frac{\gamma_k}{\gamma_0}. \tag{2.11}$$

The collection of the values of $\rho_k,\ k = 0, 1, 2, \ldots$ is called the **autocorrelation function (ACF)**. Note that by definition $\rho_0 = 1$. Also, the ACF is independent of the scale of measurement of the time series, so it is a dimensionless quantity. Furthermore, $\rho_k = \rho_{-k}$; that is, the ACF is **symmetric** around zero, so it is only necessary to compute the positive (or negative) half.

If a time series has a finite mean and autocovariance function it is said to be second-order stationary (or weakly stationary of order 2). If, in addition, the joint probability distribution of the observations at all times is multivariate normal, then that would be sufficient to result in a time series that is strictly stationary.

It is necessary to estimate the autocovariance and ACFs from a time series of finite length, say, $y_1, y_2, \ldots, y_T$. The usual estimate of the autocovariance function is

$$c_k = \hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y}), \quad k = 0, 1, 2, \ldots, K \tag{2.12}$$

and the ACF is estimated by the **sample autocorrelation function** (or **sample ACF**)

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0}, \quad k = 0, 1, \ldots, K \tag{2.13}$$

A good general rule of thumb is that at least 50 observations are required to give a reliable estimate of the ACF, and the individual sample autocorrelations should be calculated up to lag K, where K is about $T/4$.

Often we will need to determine if the autocorrelation coefficient at a particular lag is zero. This can be done by comparing the sample autocorrelation coefficient at lag k, r_k, to its standard error. If we make the assumption that the observations are uncorrelated, that is, $\rho_k = 0$ for all k, then the variance of the sample autocorrelation coefficient is

$$\mathrm{Var}(r_k) \cong \frac{1}{T} \tag{2.14}$$

and the standard error is

$$se(r_k) \cong \frac{1}{\sqrt{T}} \tag{2.15}$$

Example 2.5 Consider the chemical process viscosity readings plotted in Figure 2.9; the values are listed in Table 2.1.

The sample ACF at lag $k = 1$ is calculated as

$$\begin{aligned}
c_0 &= \frac{1}{100} \sum_{t=1}^{100-0} (y_t - \bar{y})(y_{t+0} - \bar{y}) \\
&= \frac{1}{100}[(86.7418 - 84.9153)(86.7418 - 84.9153) + \cdots \\
&\quad + (85.0572 - 84.9153)(85.0572 - 84.9153)] \\
&= 280.9332 \\
c_1 &= \frac{1}{100} \sum_{t=1}^{100-1} (y_t - \bar{y})(y_{t+1} - \bar{y}) \\
&= \frac{1}{100}[(86.7418 - 84.9153)(85.3195 - 84.9153) + \cdots \\
&\quad + (87.0048 - 84.9153)(85.0572 - 84.9153)] \\
&= 220.3137 \\
r_1 &= \frac{c_1}{c_0} = \frac{220.3137}{280.9332} = 0.7842
\end{aligned}$$

A plot and listing of the sample ACFs generated by Minitab for the first 25 lags are displayed in Figures 2.12 and 2.13, respectively.

TABLE 2.1 Chemical Process Viscosity Readings

Time Period	Reading	Time Period	Reading	Time Period	Reading	Time Period	Reading
1	86.7418	26	87.2397	51	85.5722	76	84.7052
2	85.3195	27	87.5219	52	83.7935	77	83.8168
3	84.7355	28	86.4992	53	84.3706	78	82.4171
4	85.1113	29	85.6050	54	83.3762	79	83.0420
5	85.1487	30	86.8293	55	84.9975	80	83.6993
6	84.4775	31	84.5004	56	84.3495	81	82.2033
7	84.6827	32	84.1844	57	85.3395	82	82.1413
8	84.6757	33	85.4563	58	86.0503	83	81.7961
9	86.3169	34	86.1511	59	84.8839	84	82.3241
10	88.0006	35	86.4142	60	85.4176	85	81.5316
11	86.2597	36	86.0498	61	84.2309	86	81.7280
12	85.8286	37	86.6642	62	83.5761	87	82.5375
13	83.7500	38	84.7289	63	84.1343	88	82.3877
14	84.4628	39	85.9523	64	82.6974	89	82.4159
15	84.6476	40	86.8473	65	83.5454	90	82.2102
16	84.5751	41	88.4250	66	86.4714	91	82.7673
17	82.2473	42	89.6481	67	86.2143	92	83.1234
18	83.3774	43	87.8566	68	87.0215	93	83.2203
19	83.5385	44	88.4997	69	86.6504	94	84.4510
20	85.1620	45	87.0622	70	85.7082	95	84.9145
21	83.7881	46	85.1973	71	86.1504	96	85.7609
22	84.0421	47	85.0767	72	85.8032	97	85.2302
23	84.1023	48	84.4362	73	85.6197	98	86.7312
24	84.8495	49	84.2112	74	84.2339	99	87.0048
25	87.6416	50	85.9952	75	83.5737	100	85.0572

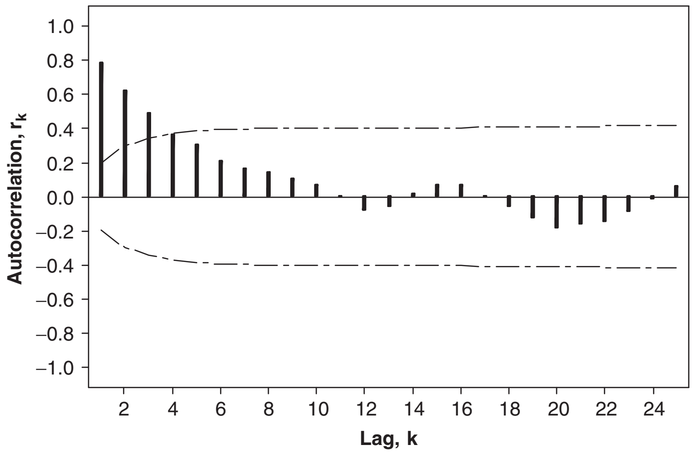

FIGURE 2.12 Sample autocorrelation function for chemical viscosity readings, with 5% significance limits.

Autocorrelation function: reading

Lag	ACF	T	LBQ
1	0.784221	7.84	63.36
2	0.628050	4.21	104.42
3	0.491587	2.83	129.83
4	0.362880	1.94	143.82
5	0.304554	1.57	153.78
6	0.208979	1.05	158.52
7	0.164320	0.82	161.48
8	0.144789	0.72	163.80
9	0.103625	0.51	165.01
10	0.066559	0.33	165.51
11	0.003949	0.02	165.51
12	−0.077226	−0.38	166.20
13	−0.051953	−0.25	166.52
14	0.020525	0.10	166.57
15	0.072784	0.36	167.21
16	0.070753	0.35	167.81
17	0.001334	0.01	167.81
18	−0.057435	−0.28	168.22
19	−0.123122	−0.60	170.13
20	−0.180546	−0.88	174.29
21	−0.162466	−0.78	177.70
22	−0.145979	−0.70	180.48
23	−0.087420	−0.42	181.50
24	−0.011579	−0.06	181.51
25	0.063170	0.30	182.06

FIGURE 2.13 Listing of sample autocorrelation functions for first 25 lags of chemical viscosity readings, Minitab session window output (the definition of T and LBQ will be given later).

Note the rate of decrease or decay in ACF values in Figure 2.12 from 0.78 to 0, followed by a sinusoidal pattern about 0. This ACF pattern is typical of stationary time series. The importance of ACF estimates exceeding the 5% significance limits will be discussed in Chapter 5. In contrast, the plot of sample ACFs for a time series of random values with constant mean has a much different appearance. The sample ACFs for pharmaceutical product sales plotted in Figure 2.14 appear randomly positive or negative, with values near zero.

While the ACF is strictly speaking defined only for a stationary time series, the sample ACF can be computed for *any* time series, so a logical question is: What does the sample ACF of a nonstationary time series look like? Consider the daily closing price for Whole Foods Market stock in Figure 1.7. The sample ACF of this time series is shown in Figure 2.15. Note that this sample ACF plot behaves quite differently than the ACF plots in Figures 2.12 and 2.14. Instead of cutting off or tailing off near zero after a few lags, this sample ACF is very **persistent**; that is, it decays very slowly and exhibits sample autocorrelations that are still rather large even at long lags. This behavior is characteristic of a nonstationary time series. Generally, if the sample ACF does not dampen out within about 15 to 20 lags, the time series is nonstationary.

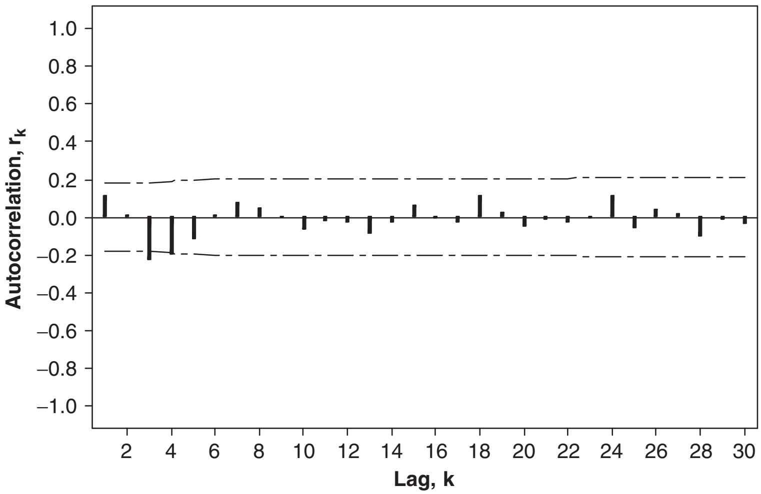

FIGURE 2.14 Autocorrelation function for pharmaceutical product sales, with 5% significance limits.

2.3.3 The Variogram

We have discussed two techniques for determining if a time series is nonstationary, plotting a reasonable long series of the data to see if it drifts or wanders away from its mean for long periods of time, and computing the sample ACF. However, often in practice there is no clear demarcation

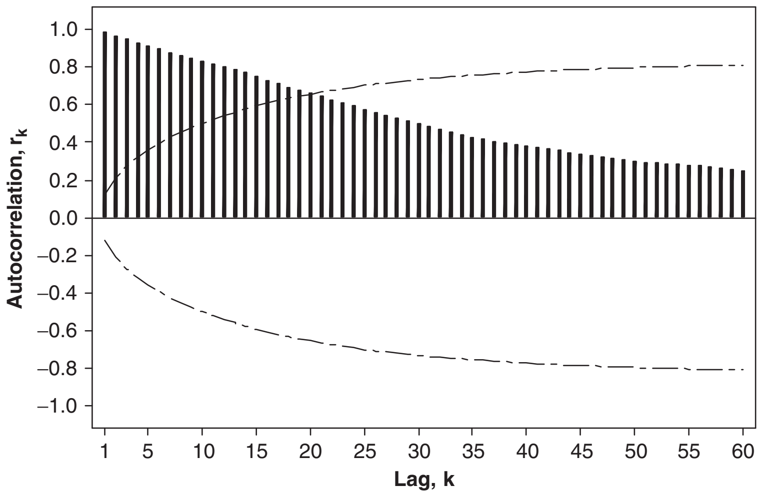

FIGURE 2.15 Autocorrelation function for Whole Foods Market stock price, with 5% significance limits.

between a stationary and a nonstationary process for many real-world time series. An additional diagnostic tool that is very useful is the **variogram**.

Suppose that the time series observations are represented by y_t. The variogram G_k measures variances of the differences between observations that are k lags apart, relative to the variance of the differences that are one time unit apart (or at lag 1). The variogram is defined mathematically as

$$G_k = \frac{\mathrm{Var}\,(y_{t+k} - y_t)}{\mathrm{Var}\,(y_{t+1} - y_t)} \quad k = 1, 2, \ldots \tag{2.16}$$

and the values of G_k are plotted as a function of the lag k. If the time series is stationary, it turns out that

$$G_k = \frac{1 - \rho_k}{1 - \rho_1},$$

but for a stationary time series $\rho_k \to 0$ as k increases, so when the variogram is plotted against lag k, G_k will reach an asymptote $1/(1 - \rho_1)$. However, if the time series is nonstationary, G_k will increase monotonically.

Estimating the variogram is accomplished by simply applying the usual sample variance to the differences, taking care to account for the changing sample sizes when the differences are taken (see Haslett (1997)). Let

$$d_t^k = y_{t+k} - y_t$$
$$\bar{d}^k = \frac{1}{T-k} \sum d_t^k.$$

Then an estimate of $\mathrm{Var}\,(y_{t+k} - y_t)$ is

$$s_k^2 = \frac{\sum_{t=1}^{T-k} \left(d_t^k - \bar{d}^k\right)^2}{T - k - 1}.$$

Therefore the sample variogram is given by

$$\hat{G}_k = \frac{s_k^2}{s_1^2} \quad k = 1, 2, \ldots \tag{2.17}$$

To illustrate the use of the variogram, consider the chemical process viscosity data plotted in Figure 2.9. Both the data plot and the sample ACF in

Lag	Variogram	Plot Variogram
1	1.0000	
2	1.7238	
3	2.3562	
4	2.9527	
5	3.2230	
6	3.6659	
7	3.8729	
8	3.9634	
9	4.1541	
10	4.3259	
11	4.6161	
12	4.9923	
13	4.8752	
14	4.5393	
15	4.2971	
16	4.3065	
17	4.6282	
18	4.9006	
19	5.2050	
20	5.4711	
21	5.3873	
22	5.3109	
23	5.0395	
24	4.6880	
25	4.3416	

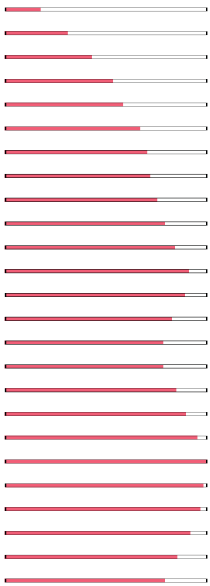

FIGURE 2.16 JMP output for the sample variogram of the chemical process viscosity data from Figure 2.19.

Figures 2.12 and 2.13 suggest that the time series is stationary. Figure 2.16 is the variogram. Many software packages do not offer the variogram as a standard pull-down menu selection, but the JMP package does. Without software, it is still fairly easy to compute.

Start by computing the successive differences of the time series for a number of lags and then find their sample variances. The ratios of these sample variances to the sample variance of the first differences will produce the sample variogram. The JMP calculations of the sample variogram are shown in Figure 2.16 and a plot is given in Figure 2.17. Notice that the sample variogram generally converges to a stable level and then fluctuates around it. This is consistent with a stationary time series, and it provides additional evidence that the chemical process viscosity data are stationary.

Now let us see what the sample variogram looks like for a nonstationary time series. The Whole Foods Market stock price data from Appendix Table B.7 originally shown in Figure 1.7 are apparently nonstationary, as it wanders about with no obvious fixed level. The sample ACF in Figure 2.15 decays very slowly and as noted previously, gives the impression that the time series is nonstationary. The calculations for the variogram from JMP are shown in Figure 2.18 and the variogram is plotted in Figure 2.19.

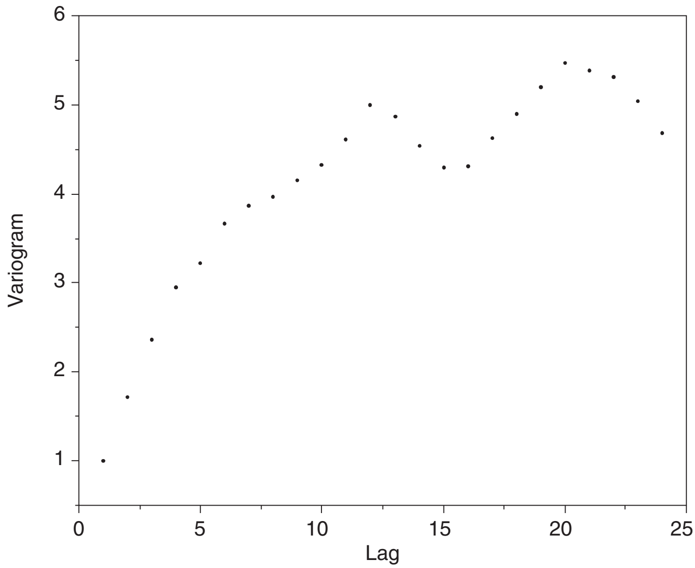

FIGURE 2.17 JMP sample variogram of the chemical process viscosity data from Figure 2.9.

Lag	Variogram	Plot Variogram
1	1.0000	
2	2.0994	
3	3.2106	
4	4.3960	
5	5.4982	
6	6.5810	
7	7.5690	
8	8.5332	
9	9.4704	
10	10.4419	
11	11.4154	
12	12.3452	
13	13.3759	
14	14.4411	
15	15.6184	
16	16.9601	
17	18.2442	
18	19.3782	
19	20.3934	
20	21.3618	
21	22.4010	
22	23.4788	
23	24.5450	
24	25.5906	
25	26.6620	

FIGURE 2.18 JMP output for the sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

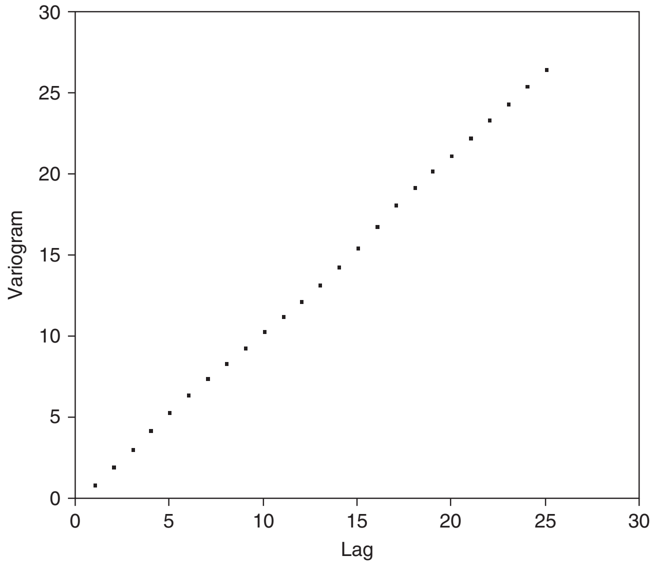

FIGURE 2.19 Sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

Notice that the sample variogram in Figure 2.19 increases monotonically for all 25 lags. This is a strong indication that the time series is nonstationary.

2.4 USE OF DATA TRANSFORMATIONS AND ADJUSTMENTS

2.4.1 Transformations

Data transformations are useful in many aspects of statistical work, often for stabilizing the variance of the data. Nonconstant variance is quite common in time series data. For example, the International Sunspot Numbers plotted in Figure 2.20a show cyclic patterns of varying magnitudes. The variability from about 1800 to 1830 is smaller than that from about 1830 to 1880; other small periods of constant, but different, variances can also be identified.

A very popular type of data transformation to deal with nonconstant variance is the **power family** of transformations, given by

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases}, \tag{2.18}$$

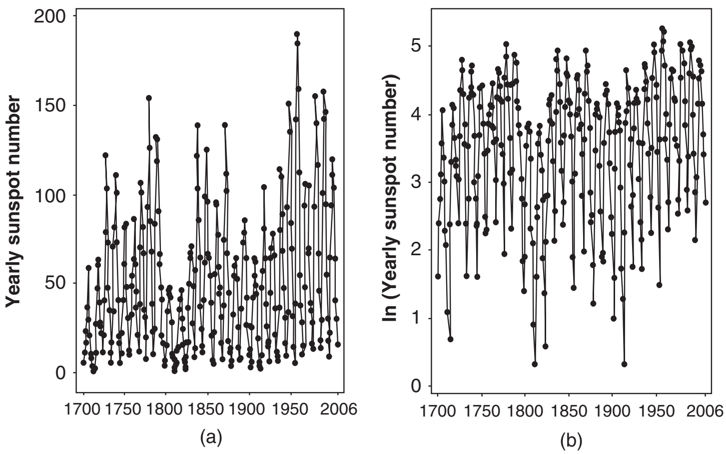

FIGURE 2.20 Yearly International Sunspot Number, (a) untransformed and (b) natural logarithm transformation. *Source*: SIDC.

where $\dot{y} = \exp[(1/T)\sum_{t=1}^{T} \ln y_t]$ is the geometric mean of the observations. If $\lambda = 1$, there is no transformation. Typical values of λ used with time series data are $\lambda = 0.5$ (a square root transformation), $\lambda = 0$ (the log transformation), $\lambda = -0.5$ (reciprocal square root transformation), and $\lambda = -1$ (inverse transformation). The divisor $\dot{y}^{\lambda-1}$ is simply a scale factor that ensures that when different models are fit to investigate the utility of different transformations (values of λ), the residual sum of squares for these models can be meaningfully compared. The reason that $\lambda = 0$ implies a log transformation is that $(y^\lambda - 1)/\lambda$ approaches the log of y as λ approaches zero. Often an appropriate value of λ is chosen empirically by fitting a model to $y^{(\lambda)}$ for various values of λ and then selecting the transformation that produces the minimum residual sum of squares.

The log transformation is used frequently in situations where the variability in the original time series increases with the average level of the series. When the standard deviation of the original series increases linearly with the mean, the log transformation is in fact an optimal variance-stabilizing transformation. The log transformation also has a very nice physical interpretation as percentage change. To illustrate this, let the time series be $y_1, y_2, \ldots, y_T$ and suppose that we are interested in the percentage change in y_t, say,

$$x_t = \frac{100(y_t - y_{t-1})}{y_{t-1}},$$

The approximate percentage change in y_t can be calculated from the differences of the log-transformed time series $x_t \cong 100[\ln(y_t) - \ln(y_{t-1})]$ because

$$100[\ln(y_t) - \ln(y_{t-1})] = 100\ln\left(\frac{y_t}{y_{t-1}}\right) = 100\ln\left(\frac{y_{t-1} + (y_t - y_{t-1})}{y_{t-1}}\right)$$
$$= 100\ln\left(1 + \frac{x_t}{100}\right) \cong x_t$$

since $\ln(1 + z) \cong z$ when z is small.

The application of a natural logarithm transformation to the International Sunspot Number, as shown in Figure 2.20b, tends to stabilize the variance and leaves just a few unusual values.

2.4.2 Trend and Seasonal Adjustments

In addition to transformations, there are also several types of adjustments that are useful in time series modeling and forecasting. Two of the most widely used are **trend adjustments** and **seasonal adjustments**. Sometimes these procedures are called trend and seasonal decomposition.

A time series that exhibits a trend is a **nonstationary** time series. Modeling and forecasting of such a time series is greatly simplified if we can eliminate the trend. One way to do this is to fit a **regression model** describing the trend component to the data and then subtracting it out of the original observations, leaving a set of residuals that are free of trend. The trend models that are usually considered are the linear trend, in which the mean of y_t is expected to change linearly with time as in

$$E(y_t) = \beta_0 + \beta_1 t \tag{2.19}$$

or as a quadratic function of time

$$E(y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 \tag{2.20}$$

or even possibly as an exponential function of time such as

$$E(y_t) = \beta_0 e^{\beta_1 t}. \tag{2.21}$$

The models in Eqs. (2.19)–(2.21) are usually fit to the data by using ordinary least squares.

Example 2.6 We will show how least squares can be used to fit regression models in Chapter 3. However, it would be useful at this point to illustrate how trend adjustment works. Minitab can be used to perform trend adjustment. Consider the annual US production of blue and gorgonzola cheeses

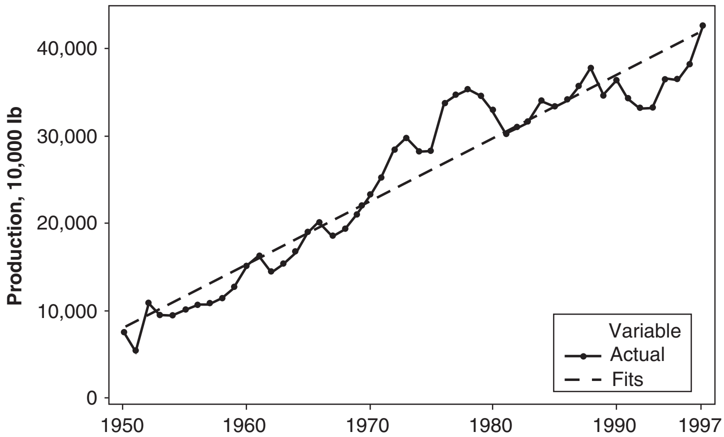

FIGURE 2.21 Blue and gorgonzola cheese production, with fitted regression line. *Source*: USDA–NASS.

shown in Figure 1.4. There is clearly a positive, nearly linear trend. The trend analysis plot in Figure 2.21 shows the original time series with the fitted line.

Plots of the residuals from this model indicate that, in addition to an underlying trend, there is additional structure. The normal probability plot (Figure 2.22a) and histogram (Figure 2.22c) indicate the residuals are

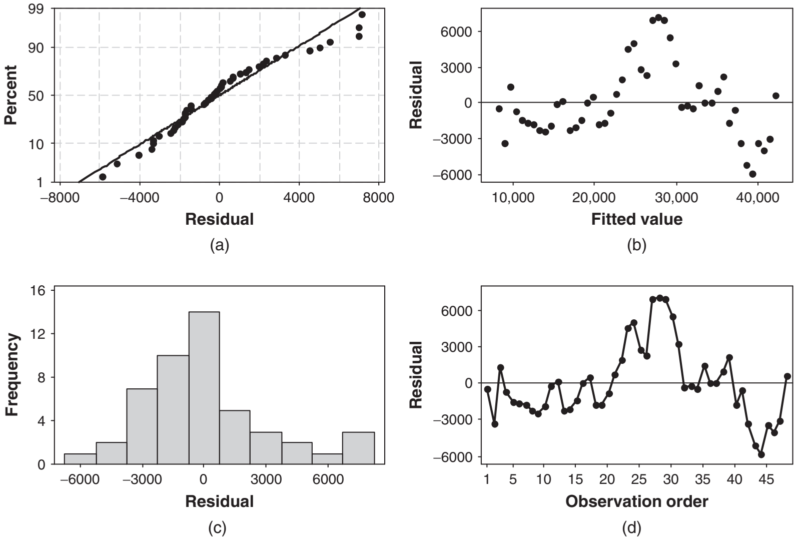

FIGURE 2.22 Residual plots for simple linear regression model of blue and gorgonzola cheese production.

approximately normally distributed. However, the plots of residuals versus fitted values (Figure 2.22b) and versus observation order (Figure 2.22d) indicate nonconstant variance in the last half of the time series. Analysis of model residuals is discussed more fully in Chapter 3.

Another approach to removing trend is by **differencing** the data; that is, applying the difference operator to the original time series to obtain a new time series, say,

$$x_t = y_t - y_{t-1} = \nabla y_t, \tag{2.22}$$

where ∇ is the (backward) difference operator. Another way to write the differencing operation is in terms of a **backshift operator** B, defined as $By_t = y_{t-1}$, so

$$x_t = (1 - B)y_t = \nabla y_t = y_t - y_{t-1} \tag{2.23}$$

with $\nabla = (1 - B)$. Differencing can be performed successively if necessary until the trend is removed; for example, the second difference is

$$x_t = \nabla^2 y_t = \nabla(\nabla y_t) = (1 - B)^2 y_t = (1 - 2B + B^2) = y_t - 2y_{t-1} + y_{t-2} \tag{2.24}$$

In general, powers of the backshift operator and the backward difference operator are defined as

$$\begin{aligned} B^d y_t &= y_{t-d} \\ \nabla^d &= (1 - B)^d \end{aligned} \tag{2.25}$$

Differencing has two advantages relative to fitting a trend model to the data. First, it does not require estimation of any parameters, so it is a more **parsimonious** (i.e., simpler) approach; and second, model fitting assumes that the trend is fixed throughout the time series history and will remain so in the (at least immediate) future. In other words, the trend component, once estimated, is assumed to be **deterministic**. Differencing can allow the trend component to change through time. The first difference accounts for a trend that impacts the change in the mean of the time series, the second difference accounts for changes in the slope of the time series, and so forth. Usually, one or two differences are all that is required in practice to remove an underlying trend in the data.

Example 2.7 Reconsider the blue and gorgonzola cheese production data. A difference of one applied to this time series removes the increasing trend (Figure 2.23) and also improves the appearance of the residuals plotted versus fitted value and observation order when a linear model is fitted to the detrended time series (Figure 2.24). This illustrates that differencing may be a very good alternative to detrending a time series by using a regression model.

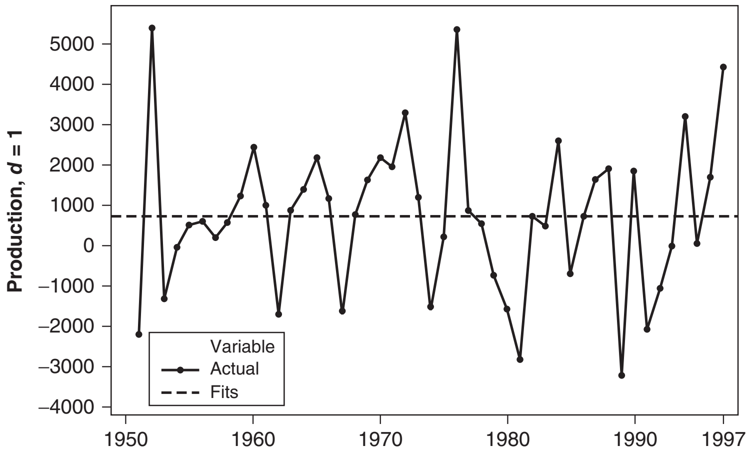

FIGURE 2.23 Blue and gorgonzola cheese production, with one difference. *Source*: USDA–NASS.

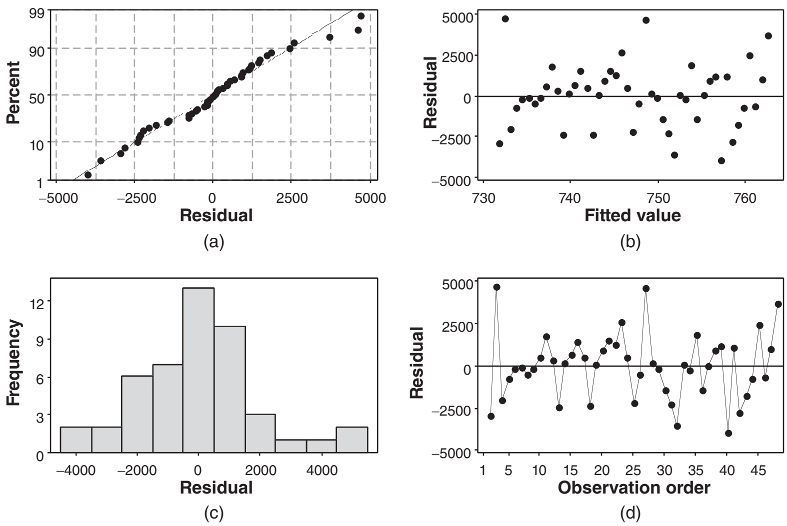

FIGURE 2.24 Residual plots for one difference of blue and gorgonzola cheese production.