Session:

Time Series and Forecasting

MODULE OBJECTIVE

At the end of this module, you will be able to:

- Explain the concept and application of forecasting and time series
- Perform time series analysis with R

SESSION OBJECTIVES

At the end of this session, you will be able to:

- Explain the nature and uses of forecasts
- Explain examples of time series
- Describe the forecasting process
- Explain data requirements for forecasting
- Use R for time series
- Use built-in functions for time series
- Perform multiple and simulated time series
- Discuss various time series models

It is difficult to make predictions, especially about the future

NEILS BOHR, Danish Physicist

Introduction to Forecasting

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.

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 session 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** shows the market yield on US Treasury Securities at 10-year constant maturity from April 1953 through December 2006. This graph is called a **time series plot**. The rate variable is collected at equally spaced time periods, as is typical in most time series and forecasting applications. Many business applications of forecasting utilize daily, weekly, monthly, quarterly, or annual data, but any reporting interval may be used. Furthermore, the data may be **instantaneous**, such as the viscosity of a chemical product at the point in time where it is measured; it may be **cumulative**, such as the total sales of a product during the month; or it may be a **statistic** that in some way reflects the activity of the variable during the time period, such as the daily closing price of a specific stock on the New York Stock Exchange.

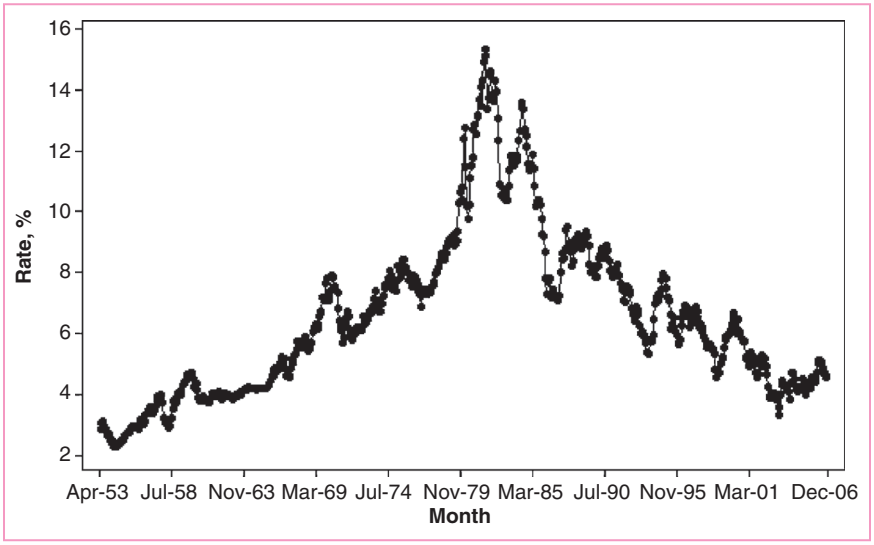


FIGURE 1
Time series plot of the market yield on US Treasury Securities at 10-year constant maturity. Source: US Treasury.

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

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 session.

Quantitative Forecasting

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.

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.

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.

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 2** appear to be constant over time, at about 10,400 × 103 units, in a random sequence with no obvious patterns. 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 average, while a value below the average tends to be followed by other values below the average.

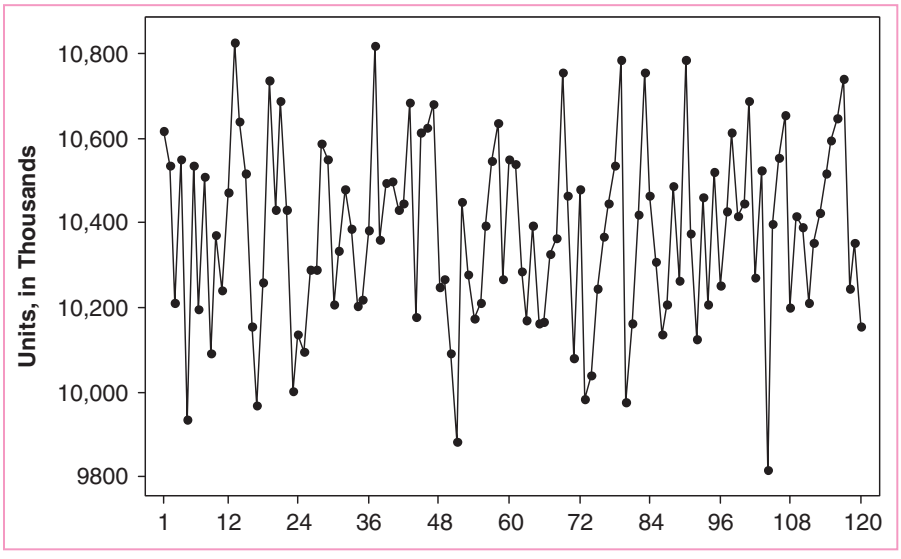


FIGURE 2
Pharmaceutical product sales.

The viscosity readings plotted in **Figure 3** exhibit autocorrelated behavior, tending to a long-run average of about 85 centipoises (cP), but with a structured, not completely random, appearance.

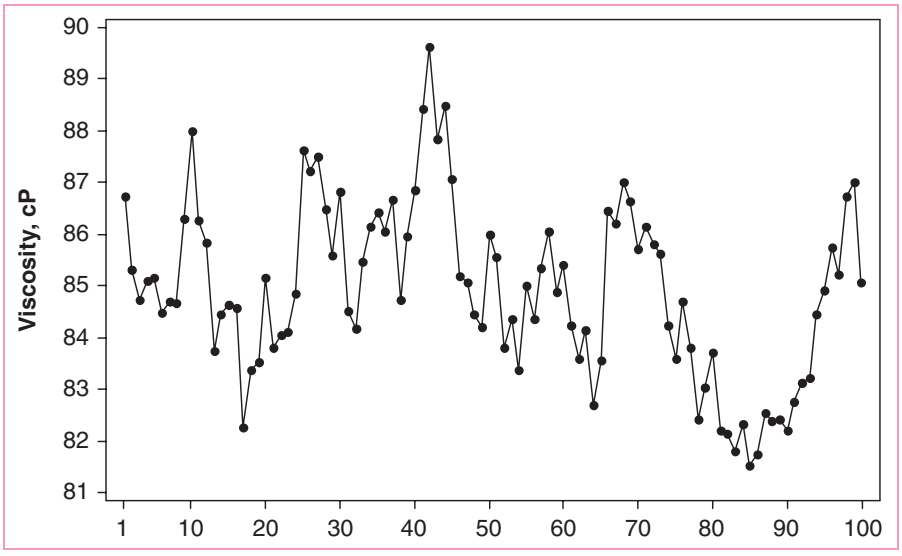


FIGURE 3
Chemical process viscosity readings.

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 103 lb) is shown in **Figure 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 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. This monthly, or seasonal, variation may be attributable to some cause such as the impact of weather on the demand for beverages.

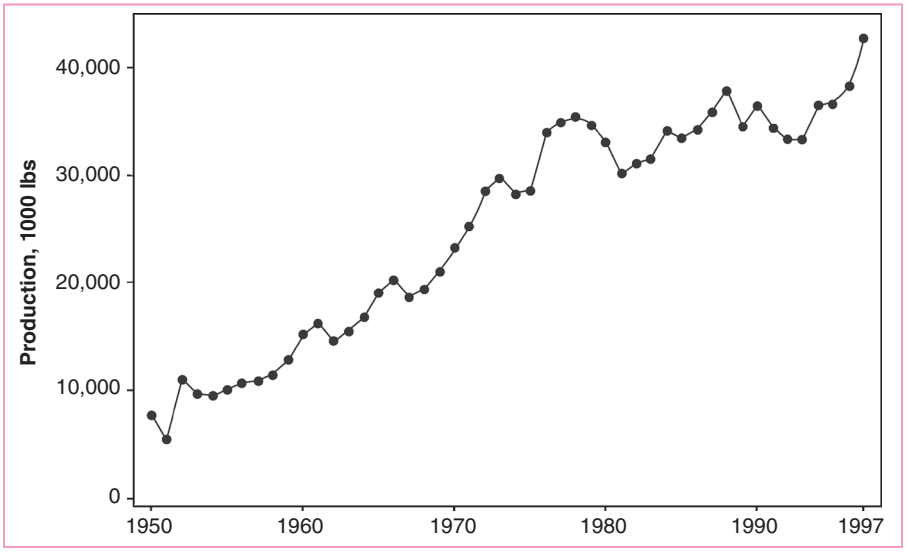


FIGURE 4
The US annual production of blue and gorgonzola cheeses. Source: USDA–NASS

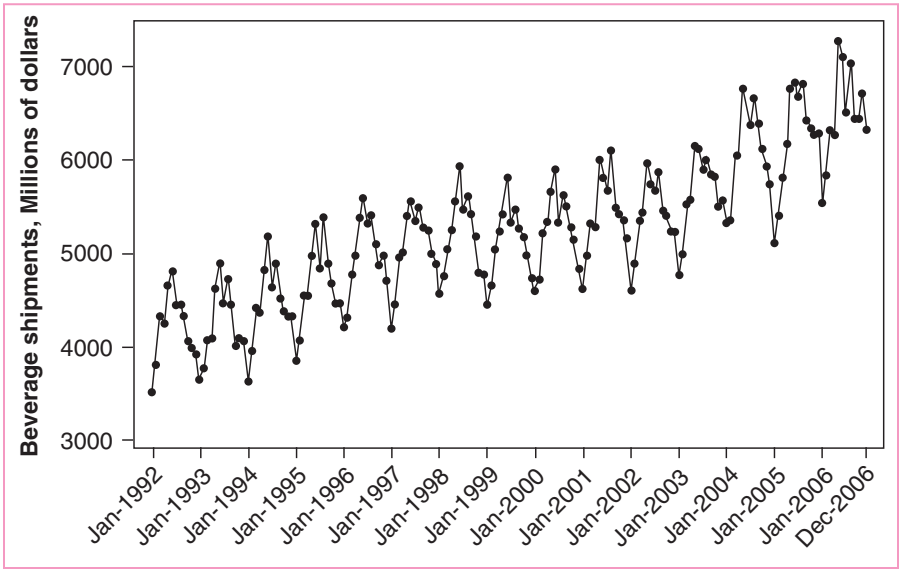


FIGURE 5
The US beverage manufacturer monthly product shipments, unadjusted. Source: US Census Bureau.

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 6**) shows an increasing trend since 1880; however, the slope, or rate of change, varies with time periods. 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 7**) exhibits a combination of patterns for both mean level and slope.

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

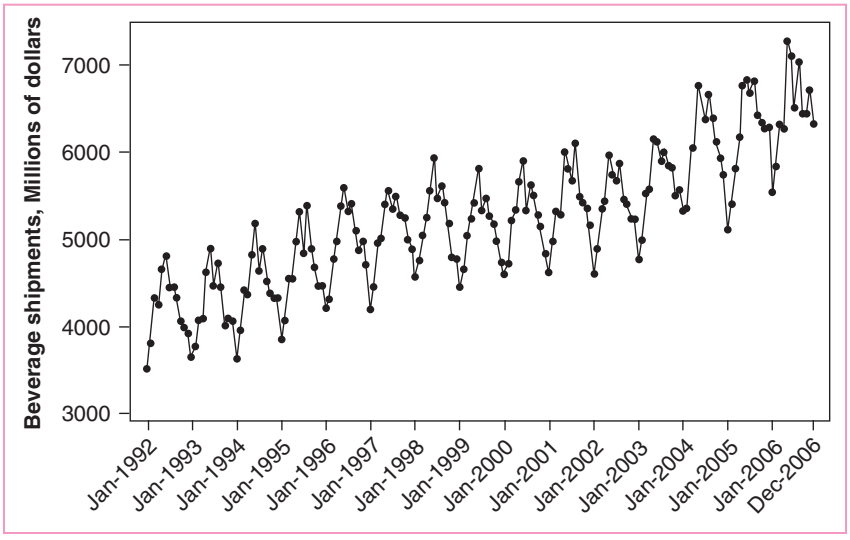


FIGURE 6
Global mean surface air temperature annual anomaly. Source: NASA-GISS.

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

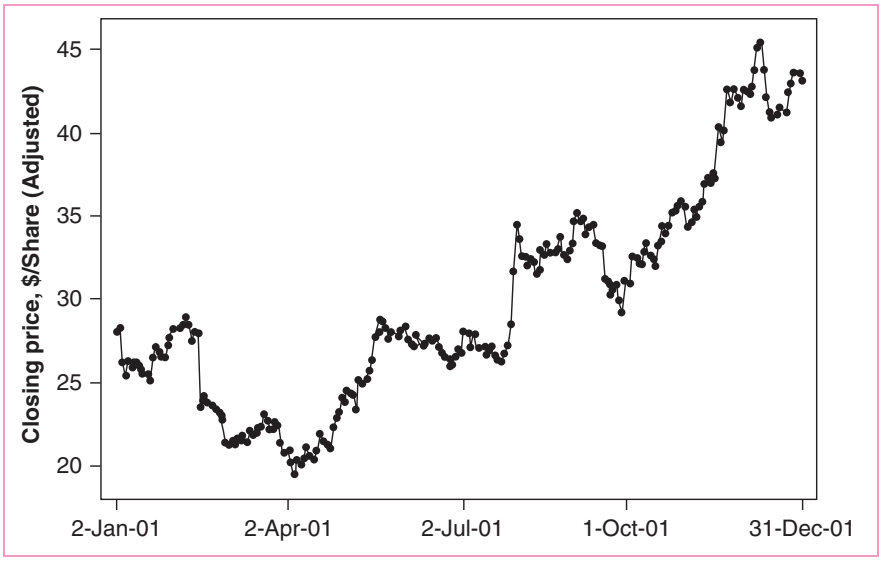


FIGURE 7
Whole foods market stock price, daily closing adjusted for splits

topics by demographic characteristics. The data are used to report on the **employment situation**, for projections with impact on hiring and training, and for a multitude of other business planning activities. The data are reported unadjusted and with seasonal adjustment to remove the effect of regular patterns that occur each year.

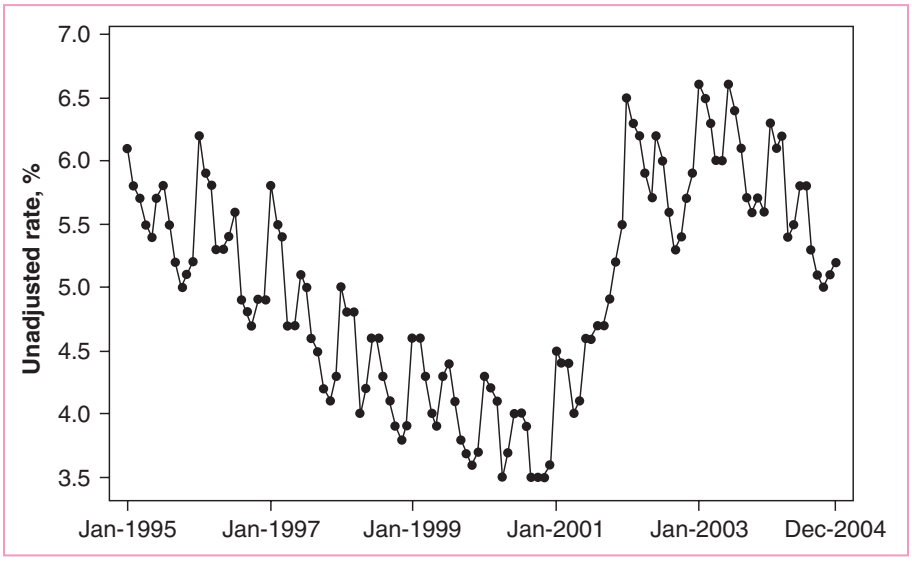


FIGURE 8
Monthly unemployment rate—full-time labor force, unadjusted. Source: US Department of Labor-BLS

The plot of monthly unadjusted unemployment rates (**Figure 8**) exhibits a mixture of patterns, similar to **Figure 5**. 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 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 9**, the plot of annual sunspot numbers reveals cyclic patterns of varying magnitudes.

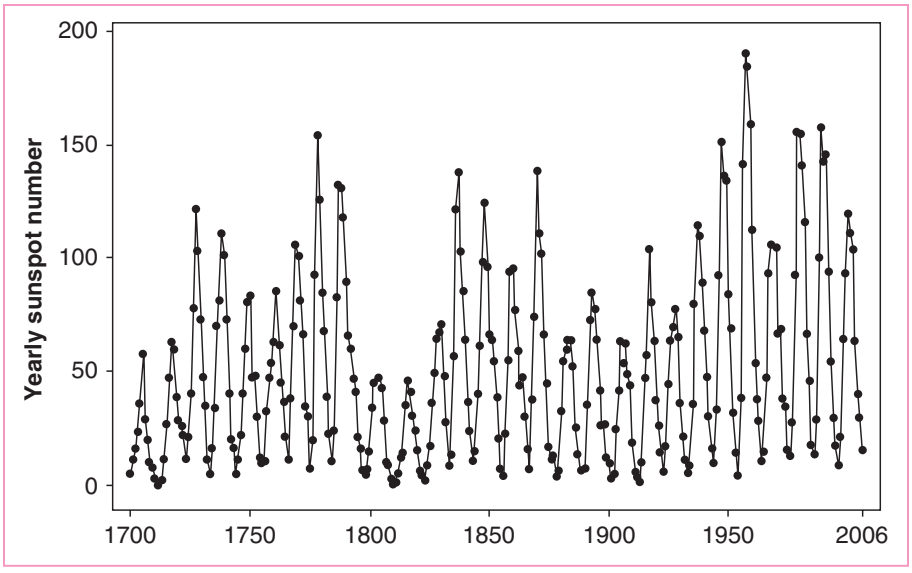


FIGURE 9
The international sunspot number. Source: SIDC.

In addition to assisting in the identification of steady-state patterns, time series plots may also draw attention to the **occurrence of atypical events**. Weekly sales of a generic pharmaceutical product dropped due to limited availability resulting from a fire at one of the four production facilities. The 5-week reduction is apparent in the time series plot of weekly sales shown in **Figure 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 11**), the measurement system was

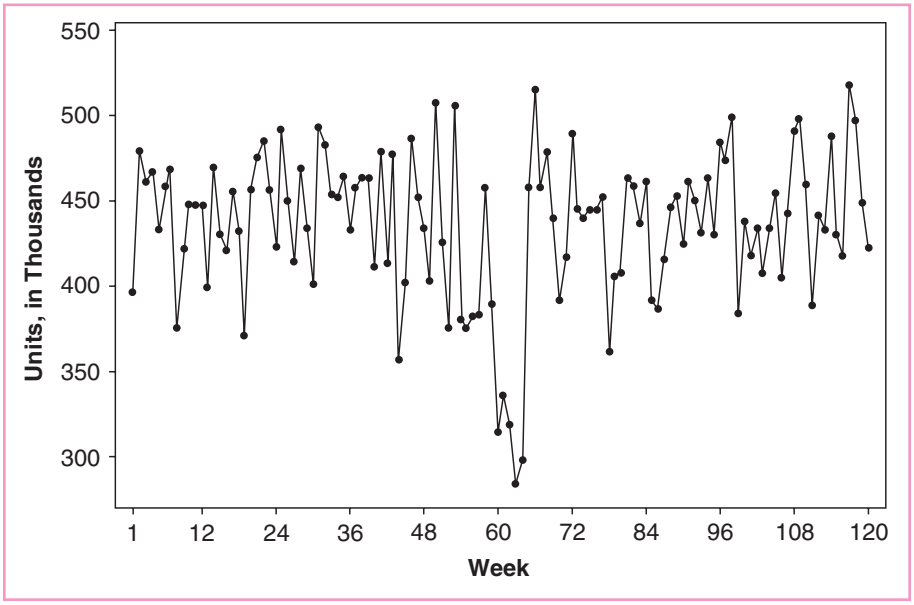


FIGURE 10
Pharmaceutical product sales.

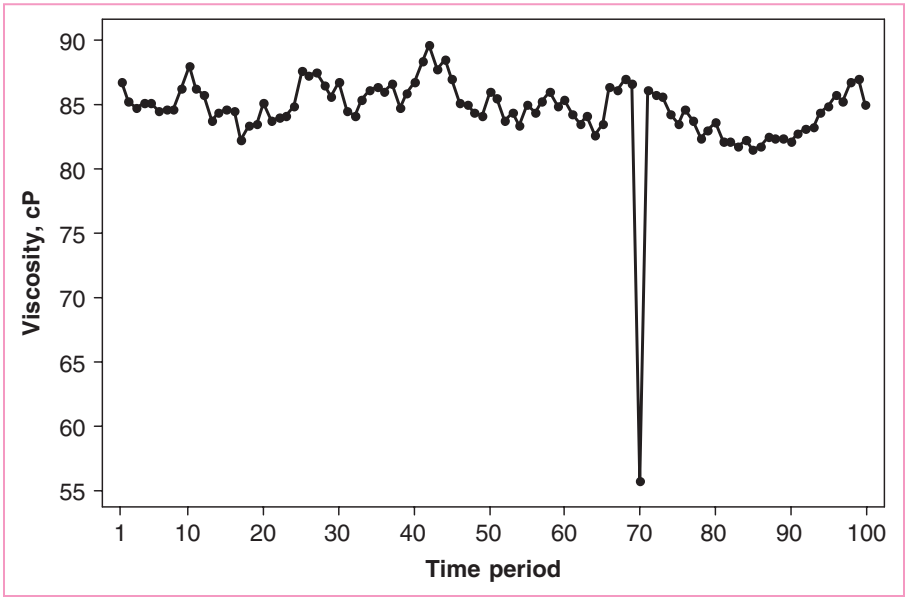


FIGURE 11
Chemical process viscosity readings, with sensor malfunction.

checked with a standard and determined to be out of calibration. The cause was determined to be a malfunctioning sensor.

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 12**.

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

meet customer demand. When multiple products are involved, the level of aggregation of the forecast (e.g., do we forecast individual products or families consisting of several similar products) can be an important consideration. Much of the ultimate success of the forecasting model in meeting the customer expectations is determined in the problem definition phase.

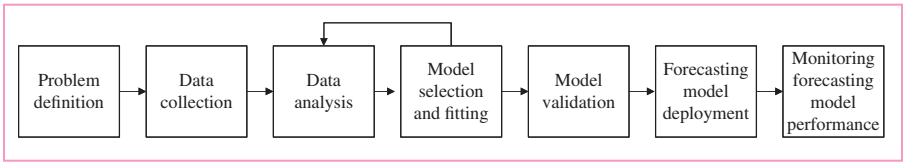


FIGURE 12
The forecasting process.

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. **Numerical summaries** of the data, such as the sample mean, standard deviation, percentiles, and autocorrelations, should also be computed and evaluated. 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 forecastingmodels and fitting the model to the data. By fitting, we mean estimating the unknown model parameters, usually by the method of least squares.

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. 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.

Data for Forecasting: 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. 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

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.

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 yj 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.

Time Series Analysis with R

Time series data are vectors of numbers, typically regularly spaced in time. Yearly counts of animals, daily prices of shares, monthly means of temperature, and minute-by-minute details of blood pressure are all examples of time series, but they are measured on different time scales. Sometimes the interest is in the time series itself (e.g. whether or not it is cyclic, or how well the data fit a particular theoretical model), and sometimes the time series is incidental to a designed experiment (e.g. repeated measures).

The three key concepts in time series analysis are

- trend,
- serial dependence, and
- stationarity.

Most time series analyses assume that the data are **untrended**. If they do showa consistent upward or downward trend, then they can be detrended before analysis (e.g. by differencing). **Serial dependence** arises because the values of adjacent members of a time series may well be correlated. **Stationarity** is a technical concept, but it can be thought of simply as meaning that the time series has the same properties wherever you start looking at it (e.g. white noise is a sequence of mutually independent random variables each with mean zero and variance $\sigma^2 > 0$).

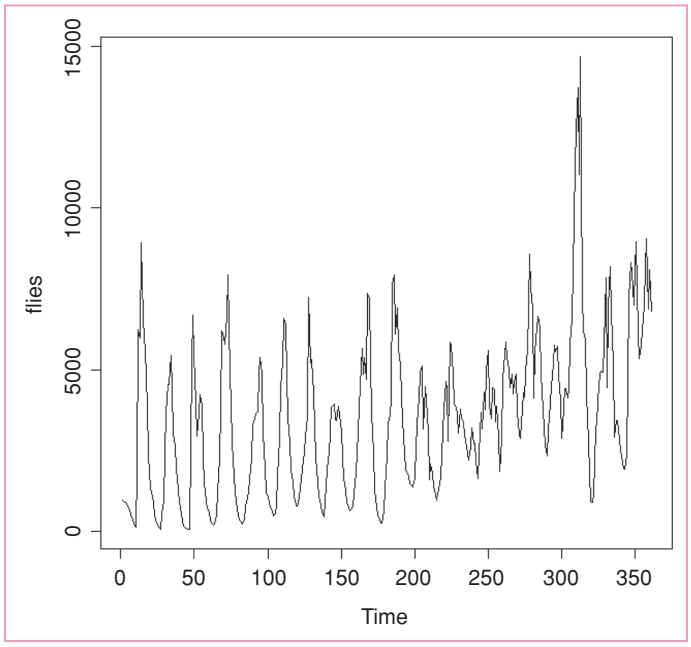
Nicholson's Blowflies

The Australian ecologist, A.J. Nicholson, reared blowfly larvae on pieces of liver in laboratory cultures that his technicians kept running continuously for almost 7 years (361 weeks, to be exact). The time series for numbers of adult flies looks like this:

```
blowfly <- read.table("c:\\temp\\blowfly.txt",header=T)
attach(blowfly)
names(blowfly)

[1] "flies"
```

First, make the flies variable into a time series object and plot it:

```
flies <- ts(flies)
plot(flies)
```



This classic time series has two clear features:

- For the first 200 weeks the system exhibits beautifully regular cycles.
- After week 200 things change (perhaps a genetic mutation had arisen); the cycles become much less clear-cut, and the population begins a pronounced upward trend.

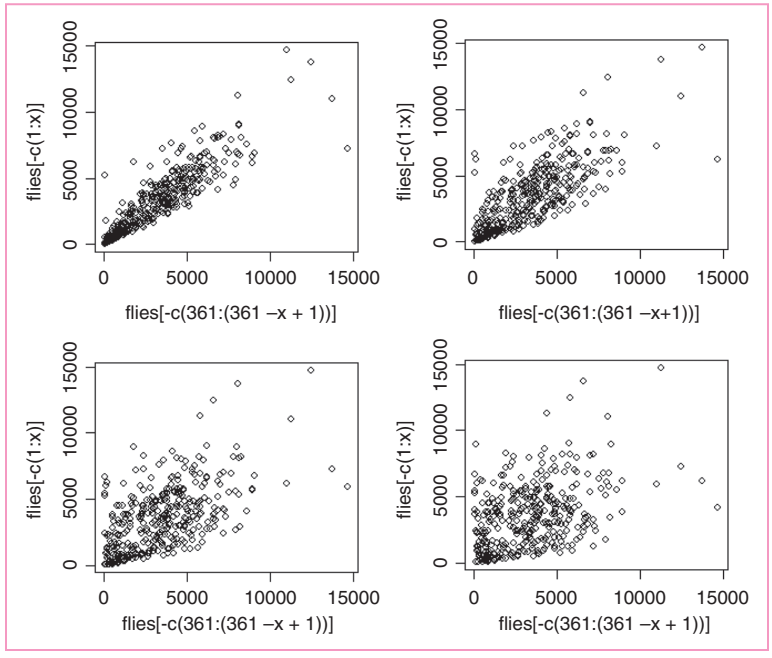
There are two important ideas to understand in time series analysis: **autocorrelation** and **partial autocorrelation**. The first describes how this week's population is related to last week's population. This is the autocorrelation at lag 1. The second describes the relationship between this week's population and the population at lag t once we have controlled for the correlations between all of the successive weeks between this week and week t. This should become clear if we draw the scatterplots from which the first four autocorrelation terms are calculated (lag 1 to lag 4).

There is a snag, however. The vector of flies at lag 1 is shorter (by one) than the original vector because the first element of the lagged vector is the second element of flies. The coordinates of the first data point to be drawn on the scatterplot are (`flies[1],flies[2]`) and the coordinates of the last plot that can be drawn are (`flies[360], flies[361]`) because the original vector is 361 element long:

```
length(flies)
[1] 361
```

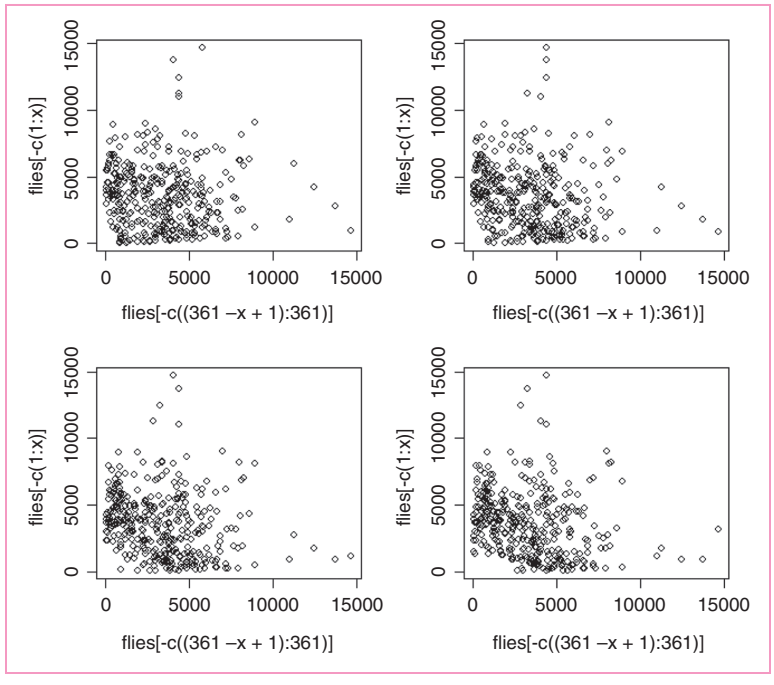
Thus, the lengths of the vectors that can be plotted go down by one for every increase in the lag of one. We can produce the four plots for lags 1 to 4 in a function like this:

```
par(mfrow=c(2,2))
sapply(1:4, function(x) plot(flies[-c(361: (361-x+1))], flies[-c(1:x)] ) )
```



The correlation is very strong at lag 1, but notice how the variance increases with population size: small populations this week are invariably correlated with small populations next week, but large populations this week may be associated with large or small populations next week. The striking pattern here is the way that the correlation fades away as the size of the lag increases. Because the population is cyclic, the correlation goes to zero, then becomes weakly negative and then becomes strongly negative. This occurs at lags that are half the cycle length. Looking back at the time series, the cycles look to be about 20 weeks in length. So let us repeat the exercise by producing scatterplots at lags of 7, 8, 9 and 10 weeks:

```
sapply(7:10, function(x) plot(flies[-c((361-x+1):361)], flies[-c(1:x)] ) )
par(mfrow=c(1,1))
```



The negative correlation at lag 10 gradually emerges from the fog of no correlation at lag 7.

More formally, the autocorrelation function $\rho(k)$ at lag k is

$$\rho(k) \quad \frac{\gamma(k)}{\gamma(A)}$$

where $\gamma(k)$ is the autocovariance function at lag k of a stationary random function $\{Y(t)\}$ given by

$$\gamma(k) = \text{cov}\{Y(t), Y(t-k)\}.$$

The most important properties of the autocorrelation coefficient are as follows:

- They are symmetric backwards and forwards, so $\rho(k) = \rho(-k)$.
- The limits are $-1 \leq \rho(k) \leq 1$.
- When $Y(t)$ and $Y(t-k)$ are independent, then $\rho(k) = 0$.
- The converse of this is not true, so that $\rho(k) = 0$ does not imply that $Y(t)$ and $Y(t-k)$ are independent
- (look at the scatterplot for $k = 7$ in the scatterplots above).

A first-order autoregressive process is written as

$$Yt = \alpha Y_{t-1} + Z_t .$$

This says that this week's population is α times last week's population plus a random term Z_t. The randomness is **white noise**; the values of Z are **serially independent**, they have a **mean of zero**, and they have **finite variance** σ^2.

In a stationary times series $-1 < \alpha < 1$. In general, then, the autocorrelation function of $\{Y(t)\}$ is

$$\rho_k = \alpha_k , \qquad k = 0, 1, 2, \ldots .$$

Partial autocorrelation is the relationship between this week's population and the population at lag t when we have controlled for the correlations between all of the successive weeks between this week and week t. That is to say, the partial autocorrelation is the correlation between $Y(t)$ and $Y(t+k)$ after regression of $Y(t)$ on $Y(t+1), Y(t+2), Y(t+3), \ldots, Y(t+k-1)$. It is obtained by solving the Yule–Walker equation

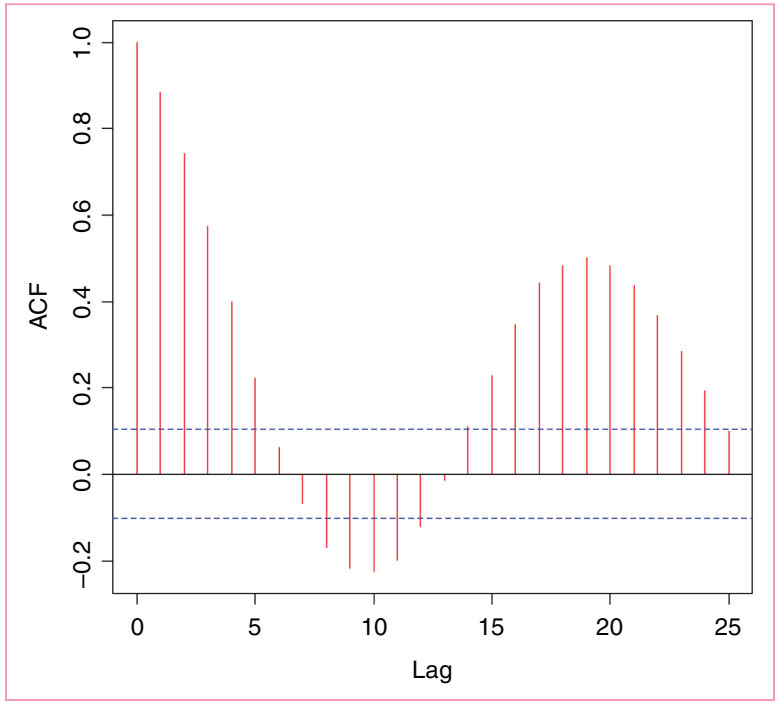
$$\rho_k = \sum_{1}^{p} \alpha_i \rho_{k-i}, \qquad k > 0,$$

with the ρ replaced by r (correlation coefficients estimated from the data). Suppose we want the partial autocorrelation between time 1 and time 3. To calculate this, we need the three ordinary correlation coefficients r_{12}, r_{13} and r_{23}. The partial $r_{13,2}$ is then

$$r_{13,2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}.$$

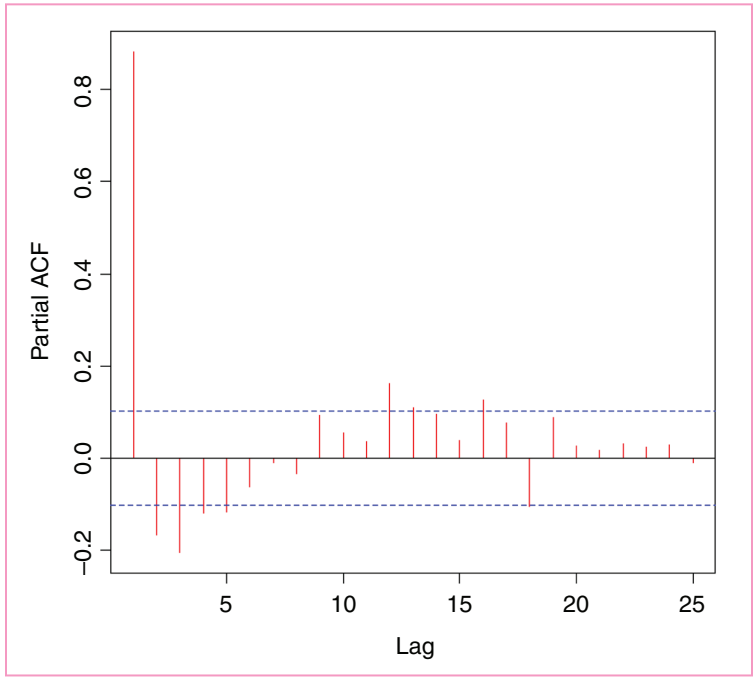
Let us look at the correlation structure of the blowfly data. The R function for calculating autocorrelations and partial autocorrelations is `acf` (the 'autocorrelation function'). First, we produce the autocorrelation plot to look for evidence of cyclic behaviour:

```
acf(flies,main="",col="red")
```



You will not see more convincing evidence of cycles than this. The blowflies exhibit highly significant, regular cycles with a period of 19 weeks. The blue dashed lines indicate the threshold values for significant correlation. What kind of time lags are involved in the generation of these cycles? We use partial autocorrelation (`type="p"`) to find this out:

```
acf(flies,type="p",main="",col="red")
```



The significant density-dependent effects are manifest at lags of 2 and 3 weeks, with other, marginally significant negative effects at lags of 4 and 5 weeks. These lags reflect the duration of the larval and pupal period (1 and 2 periods, respectively). The cycles are clearly caused by overcompensating density dependence, resulting from intraspecific competition between the larvae for food (what Nicholson christened 'scramble competition'). There is a curious positive feedback at a lag of 12 weeks (12–16 weeks, in fact). Perhaps you can think of a possible cause for this?

We should investigate the behaviour of the second half of the time series separately. Let us say it is from week 201 onwards:

```
second <- flies[201:361]
```

Now test for a linear trend in mean fly numbers against day number, from 1 to `length(second)`:

```
summary(lm(second~I(1:length(second))))
```

Note the use of `I` in the model formula (for 'as is') to tell R that the colon we have used is to generate a sequence of x values for the regression (and not an interaction term as it would otherwise have assumed).

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             2827.531    336.661     8.399 2.37e-14 ***
I(1:length(second))       21.945      3.605     6.087 8.29e-09 ***
Residual standard error: 2126 on 159 degrees of freedom
Multiple R-squared: 0.189,      Adjusted R-squared: 0.1839
F-statistic: 37.05 on 1 and 159 DF, p-value: 8.289e-09
```

This shows that there is a highly significant upward trend of about 22 extra flies on average each week in the second half of time series. We can detrend the data by subtracting the fitted values from the linear regression of `second` on day number:

```
detrended <- second - predict(lm(second~I(1:length(second))))
```

```
par(mfrow=c(2,2))
```

```
ts.plot(detrended)
```

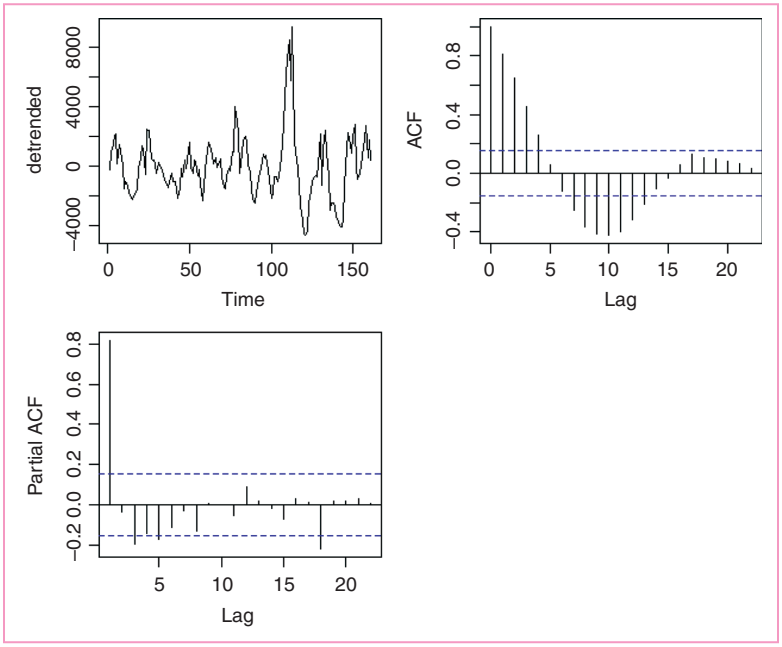
There are still cycles there, but they are weaker and less regular. We repeat the correlation analysis on the detrended data:

```
acf(detrended,main="")
```

These look more like damped oscillations than repeated cycles. What about the partials?

```
acf(detrended,type="p",main="")
```

```
par(mfrow=c(1,1))
```



There are still significant negative partial autocorrelations at lags 3 and 5, but now there is a curious extra negative partial at lag 18. It looks, therefore, as if the main features of the ecology are the same (scramble competition for food between the larvae, leading to negative partials at 3 and 5 weeks after 1 and 2 generation lags), but population size is drifting upwards and the cycles are showing a tendency to dampen out.

Moving average

The simplest way of seeing pattern in time series data is to plot the moving average. A useful summary statistic is the three-point moving average:

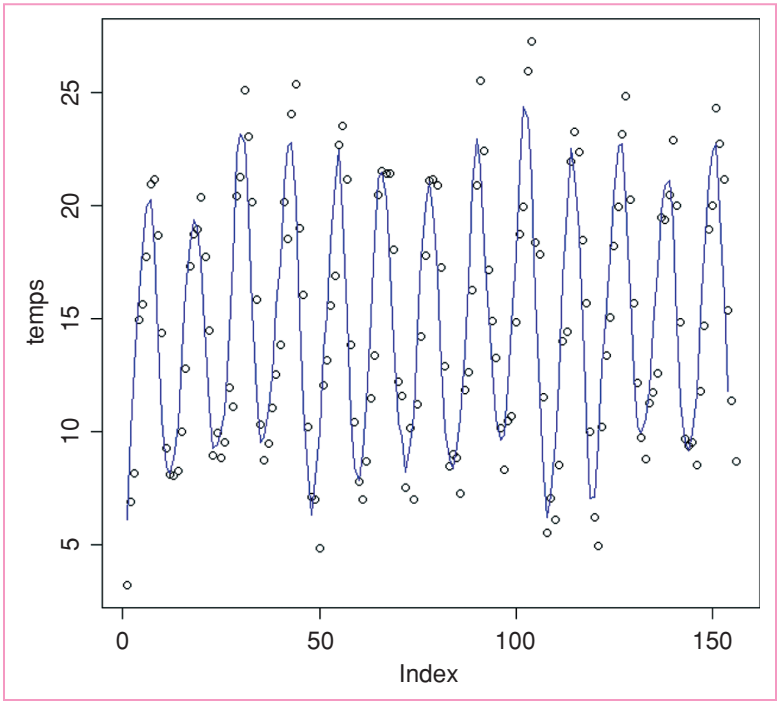
$$y_i' = \frac{y_{i-1} + y_i + y_{i+1}}{3}.$$

The function ma3 will compute the three-point moving average for any input vector x:

```
ma3 <- function (x) {
      y <- numeric(length(x)-2)
for (i in 2:(length(x)-1)) {
      y[i] <- (x[i-1]+x[i]+x[i+1])/3
}
y }
```

A time series of mean monthly temperatures will illustrate the use of the moving average:

```
temperature <- read.table("c:\\temp\\temp.txt",header=T)
attach(temperature)
tm <- ma3(temps)
plot(temps)
lines(tm[2:158],col="blue")
```



The seasonal pattern of temperature change over the 13 years of the data is clear. Note that a moving average can never capture the maxima or minima of a series (because they are averaged away). Note also that the three-point moving average is undefined for the first and last points in the series.

Seasonal data

Many time series applications involve data that exhibit seasonal cycles. The commonest applications involve weather data. Here are daily maximum and minimum temperatures from Silwood Park in south-east England over a 19-year period:

```
weather <- read.table("c:\\temp\\SilwoodWeather.txt",header=T)
attach(weather)
names(weather)
[1] "upper" "lower" "rain" "month" "yr"
plot(upper,type="l")
```

The seasonal pattern of temperature change is clear, but there is no clear trend (e.g. warming, see p. 19). Note that the x axis is labelled by the day number of the time series ('Index').

We start by modelling the seasonal component. The simplest models for cycles are scaled so that a complete annual cycle is of length 1.0 (rather than 365 days). Our series consists of 6940 days over a 19-year span, so we write:

```
length(upper)
```

```
[1] 6940
```

```
index <- 1:6940
6940/19
```

```
[1] 365.2632
```

```
time <- index/365.2632
```

The equation for the seasonal cycle is:

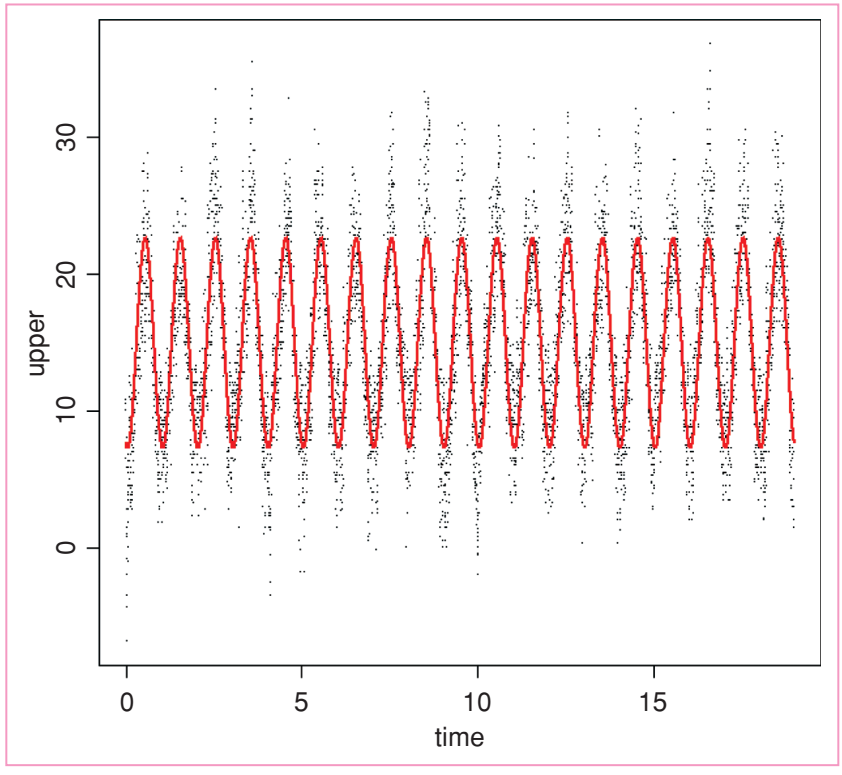
$$y = \alpha + \beta \sin(2\pi t) + \gamma \cos(2\pi t) + \varepsilon.$$

This is a linear model, so we can estimate its three parameters very simply:

```
model <- lm(upper~sin(time*2*pi)+cos(time*2*pi))
```

To investigate the fit of this model we need to plot the scattergraph using very small symbols (otherwise the fitted line will be completely obscured). The smallest useful plotting symbol is the dot "."

```
plot(time, upper, pch=".")
lines(time, predict(model),col="red",lwd=2)
```



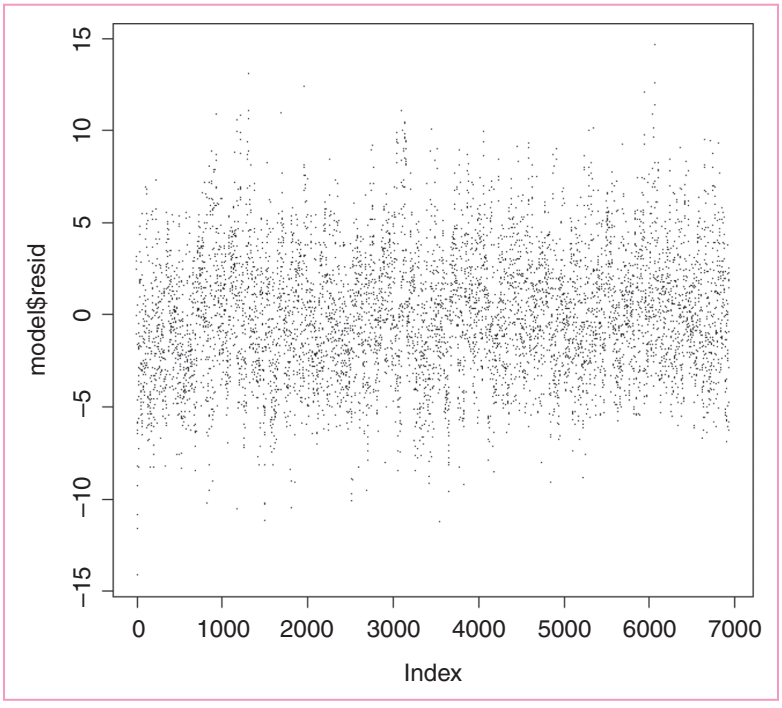
The three parameters of the model are all highly significant:

```
summary(model)

Call:

lm(formula = upper ~ sin(time * 2 * pi) + cos(time * 2 * pi))

Residuals:
     Min       1Q   Median       3Q      Max
-14.1336  -2.4220   -0.1233   2.2162   14.6456

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         14.95647   0.04088   365.86   <2e-16 ***
sin(time * 2 * pi) -2.53883   0.05781   -43.91   <2e-16 ***
```

```
cos(time * 2 * pi) -7.24017    0.05781    -125.23    <2e-16 ***

Residual standard error: 3.406 on 6937 degrees of freedom
Multiple R-squared: 0.7174,      Adjusted R-squared: 0.7173
F-statistic: 8806 on 2 and 6937 DF, p-value: < 2.2e-16
```

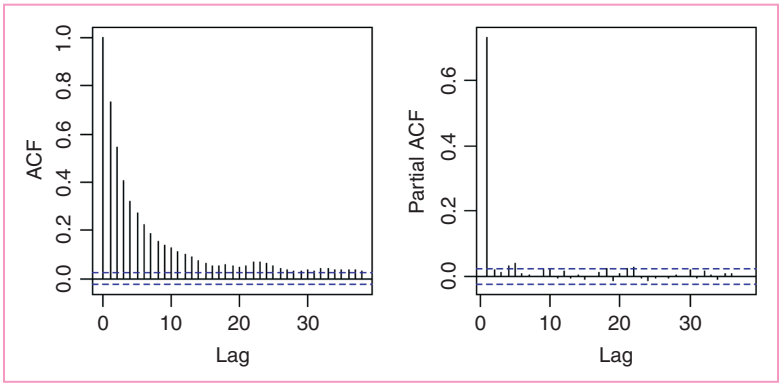
We can investigate the residuals to look for patterns (e.g. trends in the mean, or autocorrelation structure). Remember that the residuals are stored as part of the model object:

```
plot(model$resid,pch=".")
```



There looks to be some periodicity in the residuals, but no obvious trends. To look for serial correlation in the residuals, we use the `acf` function like this:

```
windows(7,4)
par(mfrow=c(1,2))
acf(model$resid,main="")
acf(model$resid,type="p",main="")
```

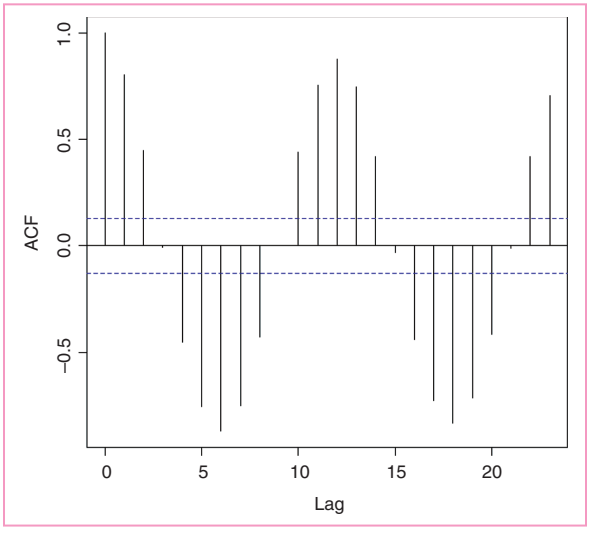


There is very strong serial correlation in the residuals, and this drops off roughly exponentially with increasing lag (left-hand graph). The partial autocorrelation at lag 1 is very large (0.7317), but the correlations at higher lags are much smaller. This suggests that an AR(1) model (autoregressive model with order 1) might be appropriate. This is the statistical justification behind the old joke about the weather forecaster who was asked what tomorrow's weather would be. 'Like today's', he said.

Pattern in the monthly means

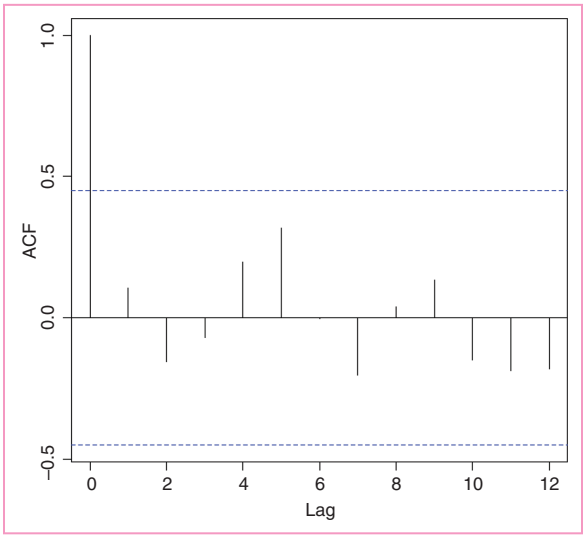
The monthly average upper temperatures show a beautiful seasonal pattern when analysed by `acf`:

```
temp <- ts(as.vector(tapply(upper,list(month,yr),mean)))
windows(7,7)
acf(temp,main="")
```



There is a perfect cycle with period 12 (as expected). What about patterns across years?

```
ytemp <- ts(as.vector(tapply(upper,yr,mean)))
acf(ytemp,main="")
```



Nothing! The pattern you may (or may not) see depends upon the scale at which you look for it. As for spatial patterns, so it is with temporal patterns. There is strong pattern between days within months (tomorrow will be like today). There is very strong pattern from month to month within years (January is cold, July is warm). But there is no pattern at all from year to year (there may be progressive global warming, but it is not apparent within this recent time series (see below), and there is absolutely no evidence for untrended serial correlation).

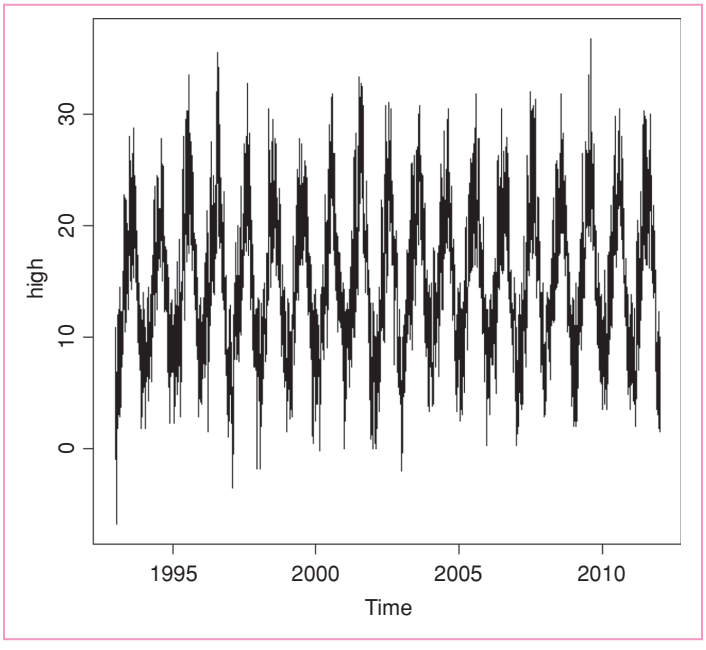
Built-in time series functions

The analysis is simpler, and the graphics are better labelled, if we convert the temperature data into a regular time series object using ts. We need to specify the first date (January 1993) as start=c(1993,1), and the number of data points per year as frequency=365.

```
high <- ts(upper,start=c(1993,1),frequency=365)
```

Now use `plot` to see a plot of the time series, correctly labelled by years:

```
plot(high)
```



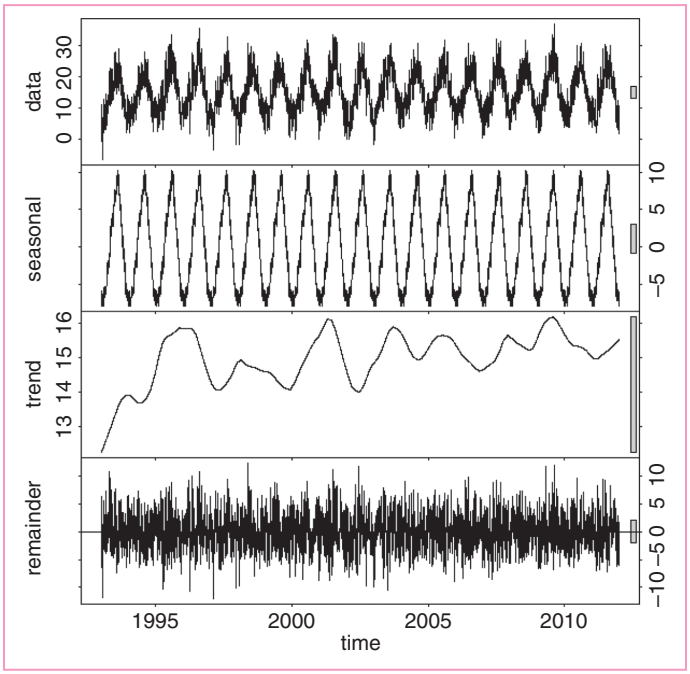
Decompositions

It is useful to be able to turn a time series into components. The function `stl` (with a lower-case letter L, not numeral one) performs seasonal decomposition of a time series into seasonal, trend and irregular components using loess. First, we make a time series object, specifying the start date and the frequency (as shown in earlier section), then use `stl` to decompose the series:

```
up <- stl(high,"periodic")
```

The `plot` function produces the data series, the seasonal component, the trend and the residuals in a single frame:

```
plot(up)
```



The remainder component is the residuals from the seasonal plus trend fit. The bars at the right-hand side are of equal heights (in user coordinates).

Testing for a Trend in the Time Series

It is important to know whether these data provide any evidence for global warming. The trend part of the figure indicates a fluctuating increase, but is it significant? The mean temperature in the last 9 years was 0.71°C higher than in the first 10 years:

```
ys <- factor(1+(yr>2002))
tapply(upper,ys,mean)
```

```
       1        2
14.62056 15.32978
```

We cannot test for a trend with linear regression because of the massive temporal pseudoreplication. Suppose we tried this:

```
model1 <- lm(upper~index+sin(time*2*pi)+cos(time*2*pi))
summary(model1)

Coefficients:
                    Estimate     Std. Error    t value   Pr(>|t|)
(Intercept)         1.433e+01    8.136e-02     176.113   <2e-16 ***
index               1.807e-04    2.031e-05     8.896     <2e-16 ***
sin(time * 2 * pi)  -2.518e+00   5.754e-02     -43.758   <2e-16 ***
cos(time * 2 * pi)  -7.240e+00   5.749e-02     -125.939  <2e-16 ***
```

```
Residual standard error: 3.387 on 6936 degrees of freedom
Multiple R-squared: 0.7206,     Adjusted R-squared: 0.7205
F-statistic: 5963 on 3 and 6936 DF, p-value: < 2.2e-16
```

It would suggest (wrongly, as we shall see) that the warming was highly significant (index p value less than 2×10^{-16} for a slope of 0.000 180 7 degrees of warming per day, leading to a predicted increase in mean temperature of 1.254°C over the 6940 days of the time series).

Since there is so much temporal pseudoreplication we should use a mixed model (`lmer`,), and because we intend to compare two models with different fixed effects we use the method of maximum likelihood (`REML=FALSE`). The explanatory variable for any trend is `index`, and we fit the model with and without this variable, allowing for different intercepts for the different years as a random effect:

```
model2 <-
lmer(upper~index+sin(time*2*pi)+cos(time*2*pi)+(1 |
factor(yr)),REML=FALSE)
model3 <-
lmer(upper~sin(time*2*pi)+cos(time*2*pi)+(1 | factor(yr)),REML=FALSE)
anova(model2,model3)
```

```
Data:
Models:
model3: upper ~ sin(time * 2 * pi) + cos(time * 2 * pi) + (1 |
factor(yr))
model2: upper ~ index + sin(time * 2 * pi) + cos(time * 2 * pi) + (1 |
model2: factor(yr))
       Df  AIC    BIC    logLik    Chisq Chi Df Pr(>Chisq)
model3 5   36452  36486  -18221
model2 6   36458  36499 -18223     0         1          1
```

Clearly, the trend is non-significant (chi-squared = 0, p = 1). If you are prepared to ignore all the variation (from day to day and from month to month), then you can get rid of the pseudoreplication by averaging and test for trend in the yearly mean values: these show a significant trend if the first year (1993) is included, but not if it is omitted:

```
means <- as.vector(tapply(upper,yr,mean))
model <- lm(means~I(1:19))
summary(model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(> | t |)
(Intercept)  14.27105  0.32220   44.293    <2e-16 ***
I(1:19)      0.06858   0.02826   2.427     0.0266 *
```

```
model <- lm(means[-1]~I(1:18))
summary(model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(> | t |)
(Intercept)  14.59826  0.30901   47.243    <2e-16 ***
I(1:18)      0.04761   0.02855   1.668     0.115
```

Obviously, you need to be circumspect when interpreting trends in time series.

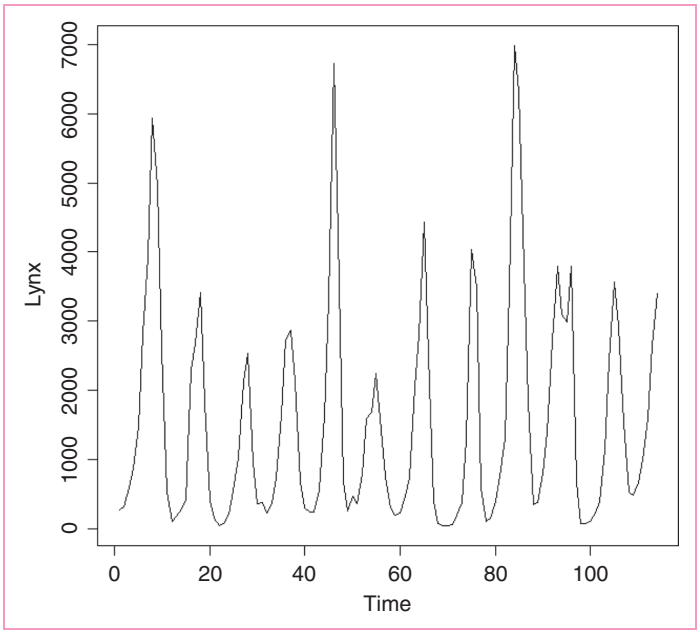
Spectral analysis

There is an alternative approach to time series analysis, which is based on the analysis of frequencies rather than fluctuations of numbers. Frequency is the reciprocal of cycle period. Ten-year cycles would have a frequency 0.1 per year. Here are the famous Canadian lynx data:

```
numbers <- read.table("c:\\temp\\lynx.txt",header=T)
attach(numbers)
names(numbers)
```

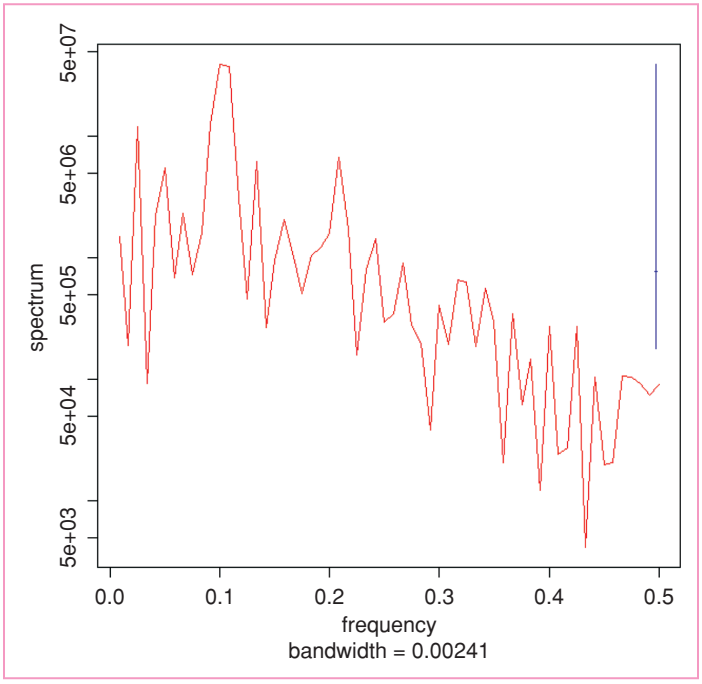
```
[1] "Lynx"
```

```
plot.ts(Lynx)
```



The fundamental tool of spectral analysis is the **periodogram**. This is based on the squared correlation between the time series and sine/cosine waves of frequency ω, and conveys exactly the same information as the autocovariance function. It may (or may not) make the information easier to interpret. Using the function is straightforward; we employ the spectrum function like this:

```
spectrum(Lynx,main="",col="red")
```



The plot is on a log scale, in units of decibels, and the subtitle on the x axis shows the bandwidth, while the 95% confidence interval in decibels is shown by the vertical blue bar in the top right-hand corner. The figure is interpreted as showing strong cycles with a frequency of about 0.1, where the maximum value of spectrum occurs. That is to say, it indicates cycles with a period of 1/0.1 = 10 years. There is a hint of longer period cycles (the local peak at frequency 0.033 would produce cycles of length 1/0.033 = 30 years) but no real suggestion of any shorter-term cycles.

Multiple Time Series

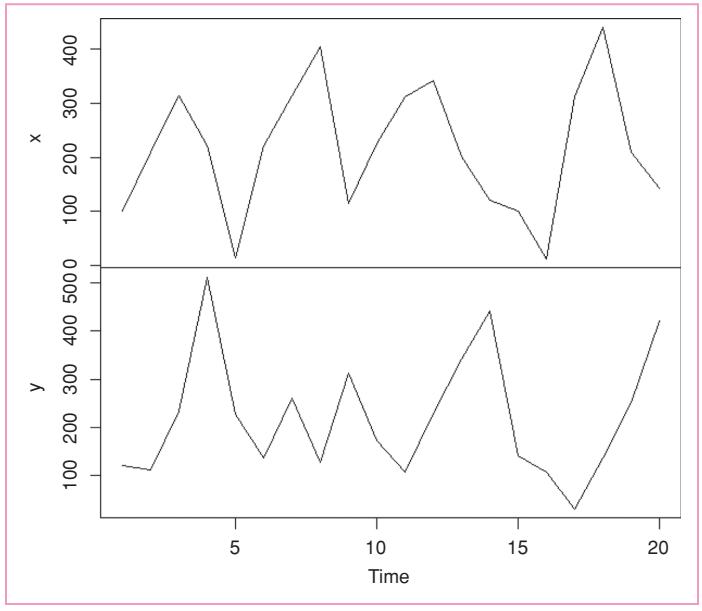
When we have two or more time series measured over the same period, the question naturally arises as to whether or not the ups and downs of the different series are correlated. It may be that we suspect that change in one of the variables causes changes in the other (e.g. changes in the number of predators may cause changes in the number of prey, because more predators means more prey eaten). We need to be careful, of course, because it will not always be obvious which way round the causal relationship might work (e.g. predator numbers may go up because prey numbers are higher; ecologists call this a numerical response). Suppose we have the following sets of counts:

```
twoseries <- read.table("c:\\temp\\twoseries.txt",header=T)
attach(twoseries)
names(twoseries)
```

```
[1] "x" "y"
```

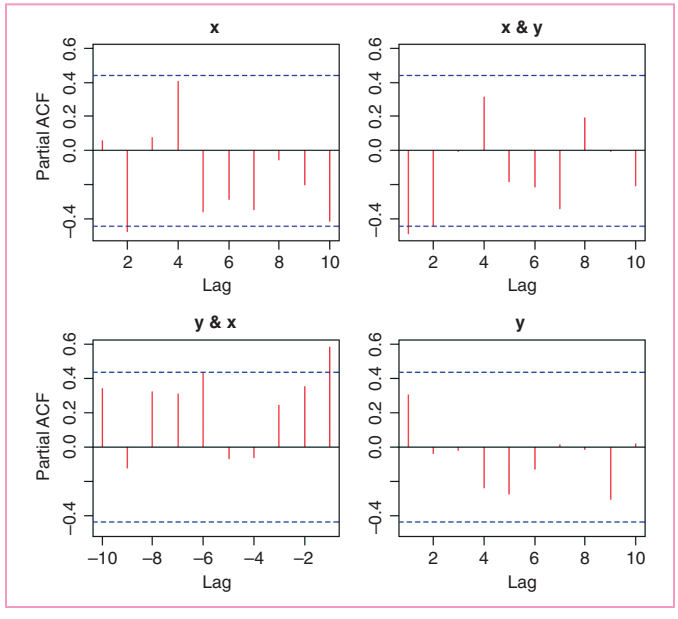
We start by inspecting the two time series one above the other:

```
plot.ts(cbind(x,y),main="")
```



There is some evidence of periodicity (at least in x) and it looks as if y lags behind x by roughly 2 periods (sometimes 1). Now let us carry out straightforward analyses on each time series separately and the crosscorrelation between the two series:

```
par(mfrow=c(2,2))
acf(cbind(x,y),type="p",col="red")
```



As we suspected, the evidence for periodicity is stronger in x than in y: the partial autocorrelation is significant and negative at lag 2 for x, but not for y. The interesting point is the cross-correlation between x and y which is significant at lags 1 and 2 (top right). Positive changes in x are associated with negative changes in y and vice versa.

Simulated time series

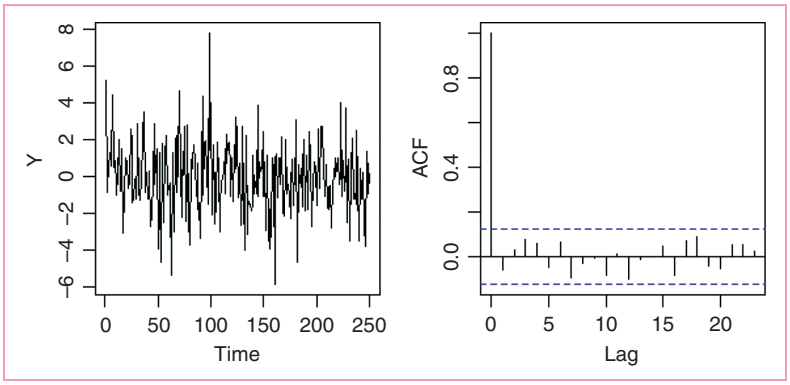
To see how the correlation structure of an AR(1) depends on the value of α, we can simulate the process over, say, 250 time periods using different values of α. We generate the white noise Z_t using the random number generator `rnorm(n,0,s)` which gives n random numbers with a mean of 0 and a standard deviation of s. To simulate the time series we evaluate

$$Y_t = \alpha Y_{t-1} + Z_t ,$$

multiplying last year's population by α then adding the relevant random number from Z_t.

We begin with the special case of $\alpha = 0$ so that $Y_t = Z_t$ and the process is pure white noise:

```
Y <- rnorm(250,0,2)
windows(7,4)
par(mfrow=c(1,2))
plot.ts(Y)
acf(Y,main="")
```



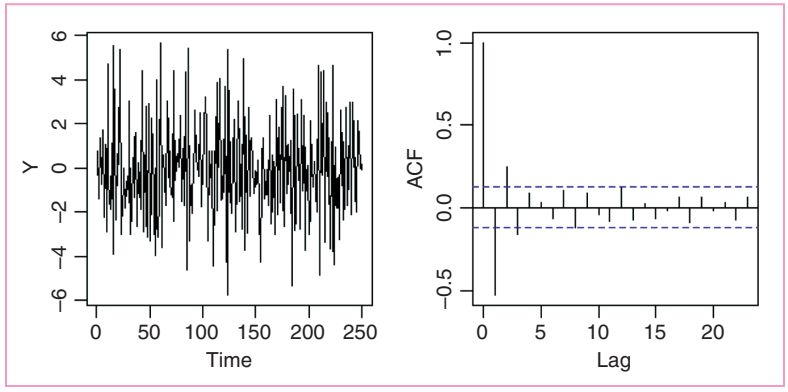
The time series is bound to be stationary because each value of Z is independent of the value before it. The correlation at lag 0 is 1 (of course), but there is absolutely no hint of any correlations at higher lags.

To generate the time series for non-zero values of α we need to use recursion: this year's population is last year's population times α plus the white noise.We begin with a negative value of $\alpha = -0.5$. First we generate all the noise values (by definition, these do not depend on population size):

```
Z <- rnorm(250,0,2)
```

Now the initial population at time 0 is set to 0 (remember that the population is stationary, so we can think of the Y values as departures from the long-term mean population size). This means that $Y_1 = Z_1$. Thus, Y_2 will be whatever Y_1 was, times −0.5, plus Z_2. And so on.

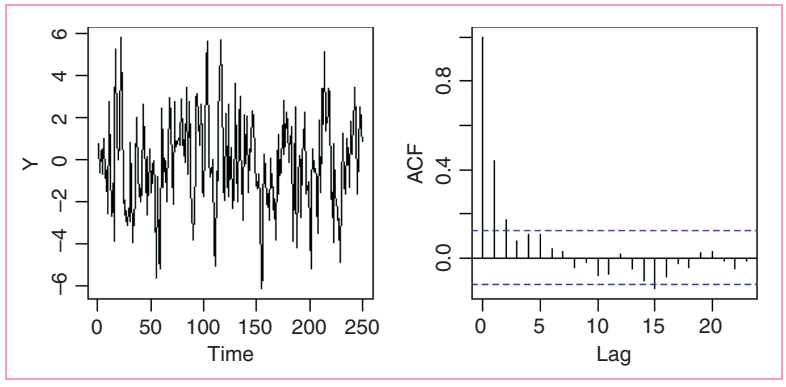
```
Y <- numeric(250)
Y[1] <- Z[1]
for (i in 2:250) Y[i] <- -0.5*Y[i-1]+Z[i]
plot.ts(Y)
acf(Y,main="")
```



The time series shows rapid return to equilibrium following random departures from it. There is a highly significant negative autocorrelation at lag 1, significant positive autocorrelation at lag 2 and so on, with the size of the correlation gradually damping away.

Let us simulate a time series with a positive value of, say, $\alpha = 0.5$:

```
Z <- rnorm(250,0,2)
Y[1] <- Z[1]
for (i in 2:250) Y[i] <- 0.5*Y[i-1]+Z[i]
plot.ts(Y)
acf(Y, main="")
```

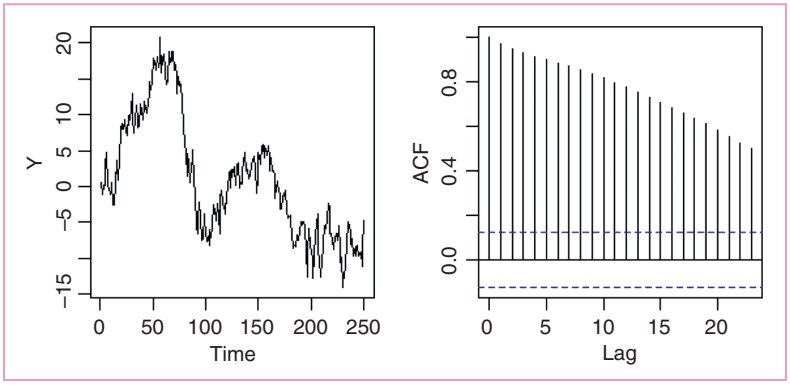


Now the time series plot looks very different, with protracted periods spent drifting away from the long-term average. The autocorrelation plot shows significant positive correlations for the first three lags.

Finally, we look at the special case of $\alpha = 1$. This means that the time series is a classic **random walk**, given by

$$Y_t = Y_{t-1} + Z_t$$

```
Z <- rnorm(250,0,2)
Y[1] <- Z[1]
for (i in 2:250) Y[i] <- Y[i-1]+Z[i]
plot.ts(Y)
acf(Y, main="")
```



The time series wanders about and strays far away from the long-term average. The `acf` plot shows positive correlations dying away very slowly, and still highly significant at lags of more than 20. Of course, if you do another realization of the process, the time series will look very different, but the autocorrelations will be similar.

Time series models

Time series models come in three kinds (Box and Jenkins, 1976):

- **moving average (MA) models** where

$$X_t = \sum_{j=0}^{q} \beta_j \varepsilon_{t-j};$$

- **autoregressive (AR) models** where

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t;$$

- **autoregressive moving average (ARMA) models** where

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=0}^{q} \beta_j \varepsilon_{t-j}.$$

A moving average of order q averages the random variation over the last q time periods. An autoregressive model of order p computes X_t as a function of the last p values of X, so, for a second-order process, we would use

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t.$$

Typically, we would use the partial autocorrelation plot (above) to determine the order. So, for the lynx data (p. 28) we would use order 2 or 4, depending on taste. Other things being equal, parsimony suggests the use of order 2. The fundamental difference is that a set of random components (ε_{t-j}) influences the current value of a MA process, whereas only the current random effect (ε_t) affects an AR process. Both kinds of effects are at work in an ARMA processes. Ecological models of population dynamics are typically AR models. For instance,

$$N_t = \lambda N_{t-1}$$

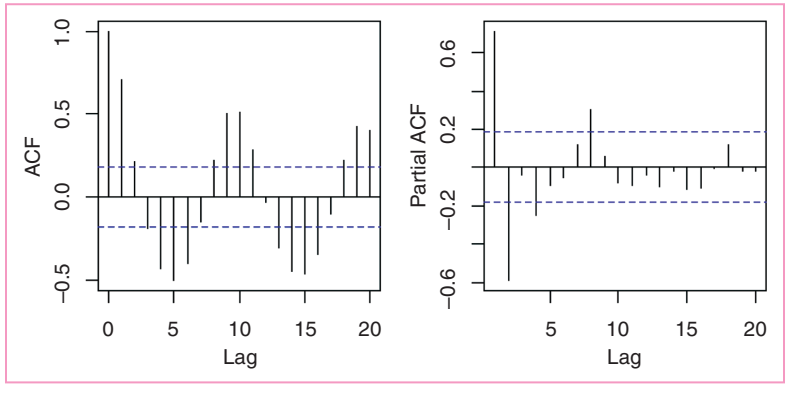
(the discrete-time version of exponential growth ($\lambda > 1$) or decay ($\lambda < 1$)) looks just like an first order AR process with the random effects missing. This is somewhat misleading, however, since time series are supposed to be

stationary, which would imply a long-term average value of $\lambda = 1$. But, in the absence of density dependence (as here), this is impossible. The α of the AR model is not the λ of the population model.

Models are fitted using the `arima` function, and their performances are compared using the AIC (see p. 415). The most important component of the model is `order`. This is a vector of length 3 specifying the order of the autoregressive operators, the number of differences, and the order of moving average operators. Thus `order=c(1,3,2)` is based on a first-order autoregressive process, three differences, and a secondorder moving average. The Canadian lynx data are used as an example of `arima` in time series modelling.

Records of the number of skins of predators (lynx) and prey (snowshoe hares) returned by trappers were collected over many years by the Hudson's Bay Company. The lynx numbers are shown on p. 28 and exhibit a clear 10-year cycle. We begin by plotting the autocorrelation and partial autocorrelation functions:

```
windows(7,4)
par(mfrow=c(1,2))
acf(Lynx,main="")
acf(Lynx,type="p",main="")
```



The population is very clearly cyclic, with a period of 10 years. The dynamics appear to be driven by strong, negative density dependence (a partial autocorrelation of –0.588) at lag 2. There are other significant partials at lag 1 and lag 8 (positive) and lag 4 (negative). Of course you cannot infer the mechanism by observing the dynamics, but the lags associated with significant negative and positive feedbacks are extremely interesting and highly suggestive. The main prey species of the lynx is the snowshoe hare and the negative feedback at lag 2 may reflect the timescale of this predator–prey interaction. The hares are known to cause medium-term induced reductions in the quality of their food plants as a result of heavy browsing pressure when the hares are at high density, and this could map through to lynx populations with lag 4.

The `order` vector specifies the non-seasonal part of the ARIMA model: the three components (p, d, q) are the AR order, the degree of differencing, and the MA order. We start by investigating the effects of AR order with no differencing and no moving average terms, comparing models on the basis of the AIC:

```
model10 <- arima(Lynx,order=c(1,0,0))
model20 <- arima(Lynx,order=c(2,0,0))
model30 <- arima(Lynx,order=c(3,0,0))
model40 <- arima(Lynx,order=c(4,0,0))
model50 <- arima(Lynx,order=c(5,0,0))
model60 <- arima(Lynx,order=c(6,0,0))
AIC(model10,model20,model30,model40,model50,model60)
          df      AIC
model10    3 1926.991
model20    4 1878.032
model30    5 1879.957
model40    6 1874.222
```

```
model150   7 1875.276
model160   8 1876.858
```

On the basis of AR alone, it appears that order 4 is best (AIC = 1874.222). What about MA?

```
model101 <- arima(Lynx,order=c(0,0,1))
model102 <- arima(Lynx,order=c(0,0,2))
model103 <- arima(Lynx,order=c(0,0,3))
model104 <- arima(Lynx,order=c(0,0,4))
model105 <- arima(Lynx,order=c(0,0,5))
model106 <- arima(Lynx,order=c(0,0,6))
AIC(model101,model102,model103,model104,model105,model106)
          df     AIC
model101   3 1917.947
model102   4 1890.061
model103   5 1887.770
model104   6 1888.279
model105   7 1885.698
model106   8 1885.230
```

The AIC values are generally higher than given by the AR models. Perhaps there is a combination of AR and MA terms that is better than either on their own?

```
model40 <- arima(Lynx,order=c(4,0,0))
model41 <- arima(Lynx,order=c(4,0,1))
model42 <- arima(Lynx,order=c(4,0,2))
model43 <- arima(Lynx,order=c(4,0,3))
AIC(model40,model41,model42,model43)
          df     AIC
model40    6 1874.222
model41    7 1875.351
model42    8 1862.435
model43    9 1880.432
```

Evidently there is no need for a moving average term (model40 is best). What about the degree of differencing?

```
model400 <- arima(Lynx,order=c(4,0,0))
model401 <- arima(Lynx,order=c(4,1,0))
model402 <- arima(Lynx,order=c(4,2,0))
model403 <- arima(Lynx,order=c(4,3,0))
AIC(model400,model401,model402,model403)
          df     AIC
model400   6 1874.222
model401   5 1890.961
model402   5 1917.882
model403   5 1946.143
```

The model with no differencing performs best. The lowest AIC is 1874.222, which suggests that a model with an AR lag of 4, no differencing and no moving average terms is best. This implies that a rather complex ecological model is required which takes account of both the significant partial correlations at lags of 2 and 4 years, and not just the 2-year lag (i.e. plant–herbivore effects may be necessary to explain the dynamics, in addition to predator–prey effects).

Cheat Sheet

- A forecast is a prediction of some future event or events.
- Forecasting is an important problem that spans many fields and forecasting problems are often classified as short-term, medium-term, and long-term.
- 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.
- Despite the wide range of problem situations that require forecasts, there are only two broad types of forecasting techniques—qualitative methods and quantitative methods.
- Time series plots can reveal patterns such as random, trends, level shifts, periods or cycles, unusual observations, or a combination of patterns.
- The activities in the forecasting process are: Problem definition; Data collection; Data analysis; Model selection and fitting; Model validation; Forecasting model deployment; Monitoring forecasting model performance
- There are two important ideas to understand in time series analysis: autocorrelation and partial autocorrelation.
- The simplest way of seeing pattern in time series data is to plot the moving average.
- When we have two or more time series measured over the same period, the question naturally arises as to whether or not the ups and downs of the different series are correlated. It may be that we suspect that change in one of the variables causes changes in the other
- Time series models can be moving average (MA), autoregressive (AR) or autoregressive moving average (ARIMA) model,