Time Series Data Analysis and Forecasting

MSDS

Module 1

Introduction to Time-series Data

Table of Content

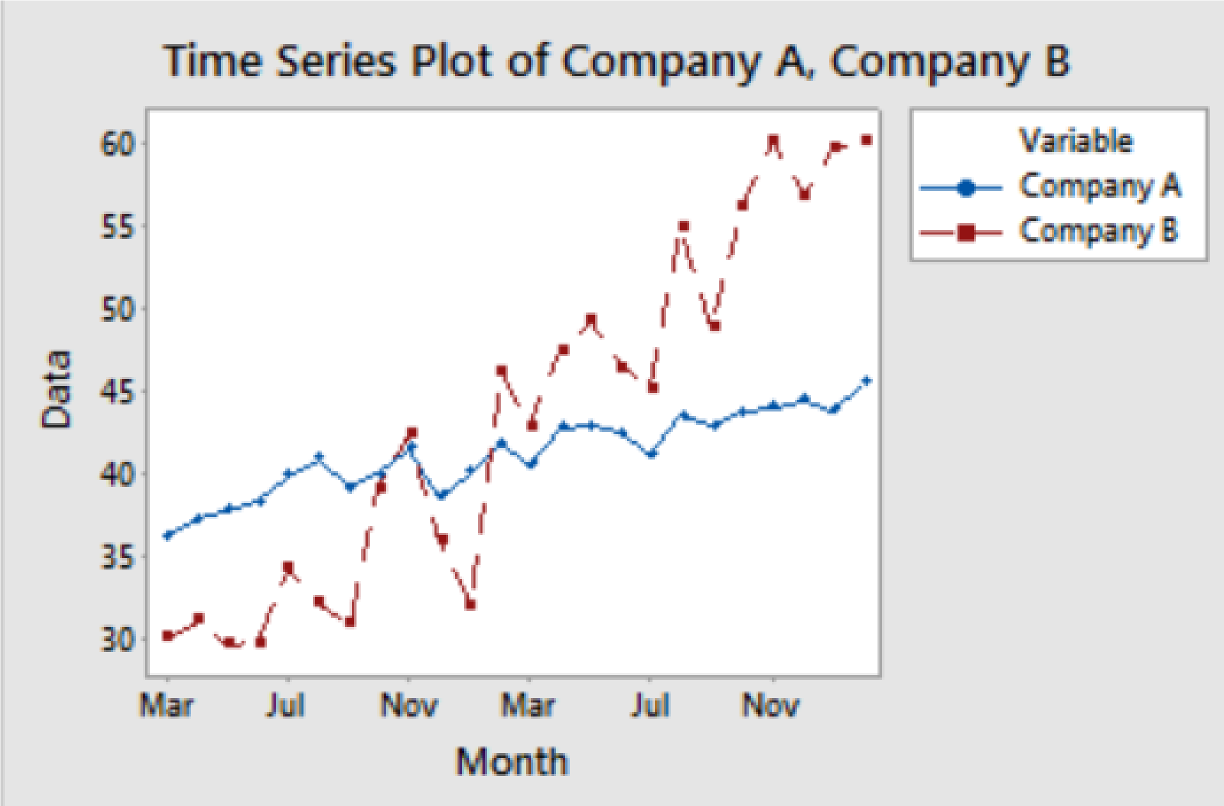
- Time Series data
- Introduction to Time series Analysis and forecasting
- Characteristics of Time series Data
- Components of time-series
- trend, seasonality, and noise
- Time-series data visualization techniques

What is Time Series Data?

- Time series data refers to a sequence of data points collected and recorded at regular time intervals.
- These intervals can be **seconds**, **minutes**, **hours**, **days**, **months**, or even **years**, depending on the context and the frequency of data collection.
- Time series data is commonly used in fields such as finance, economics, weather forecasting, signal processing.

> ➢ A time series is a series of observation x_t over a period of time
>
> ➢ Observation can be randomly sampled over an entire interval

Example : Time series data



- Company A shows a slow increase over the two-year period.
- The dashed line for Company B also shows an overall increase for the two years, but it fluctuates more than that of Company A.
- Company B starts lower than Company A, but Company B surpasses Company A by April.

Example :Johnson & Johnson Quarterly Earnings

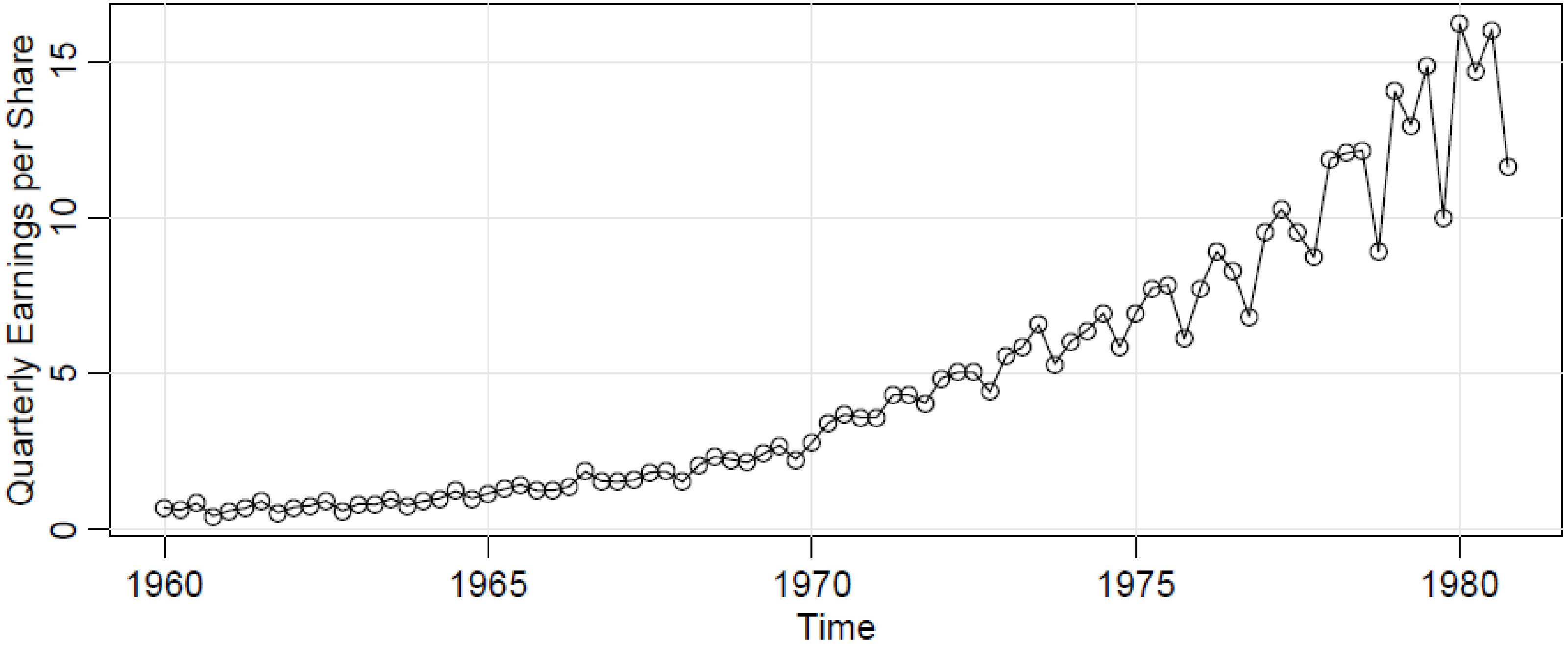


Fig. 1.1. *Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.*

Shows quarterly earnings per share for the U.S. company Johnson & Johnson, furnished by Professor Paul Griffin (personal communication) of the Graduate School of Management, University of California, Davis.

Example :**Global Warming**

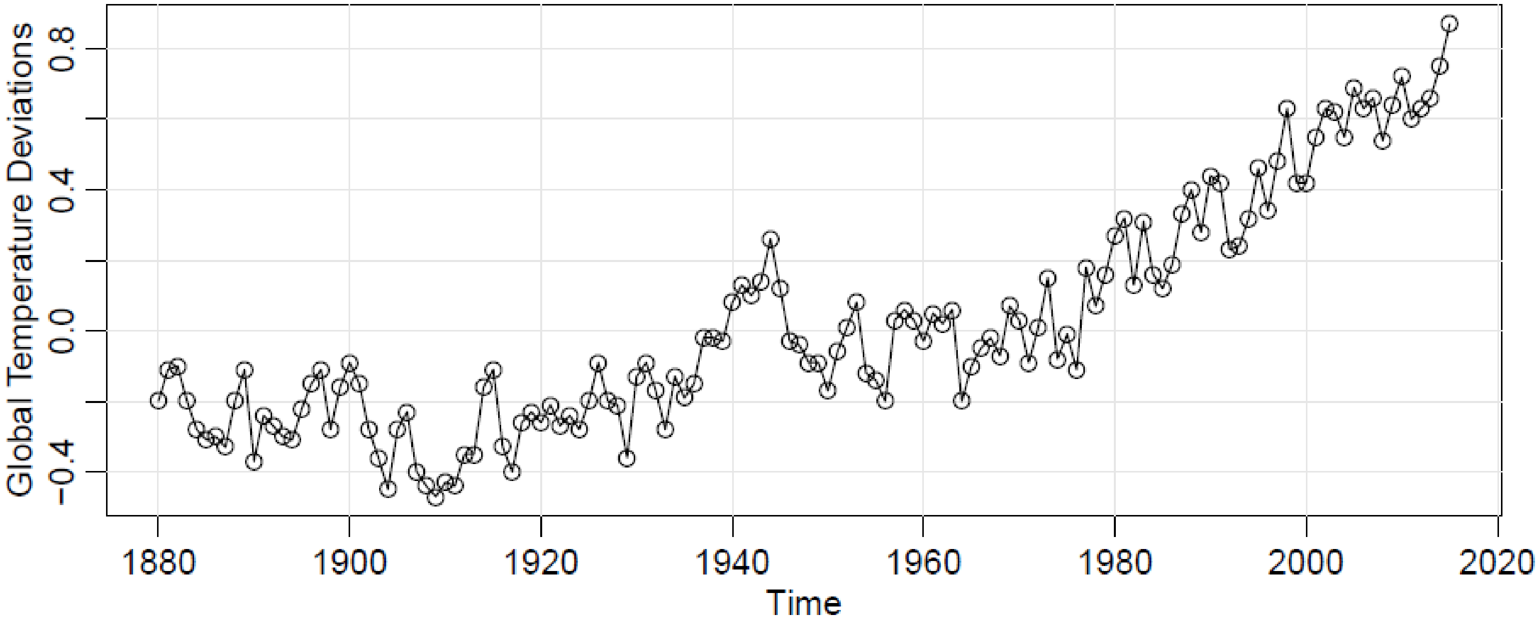


Fig. 1.2. *Yearly average global temperature deviations (1880–2015) in degrees centigrade.*

- The data are the global mean land–ocean temperature index from 1880 to 2015, with the base period 1951-1980.
- upward trend in the series during the latter part of the twentieth century which support global warming hypothesis.

Why Time series analysis and forecasting?

- time series analysis and forecasting play a vital role in decision-making processes across diverse industries
- it provides insights into past trends, current conditions, and future projections.
- These techniques help organizations and individuals make more informed and effective decisions in a dynamic and uncertain environment.

Time series analysis and forecasting??

Time series analysis and forecasting are important for ---

- **Pattern Recognition**:

  -it helps in identifying patterns, trends, and seasonality

  -By understanding historical patterns

  -businesses can make informed decisions about future strategies

  -inventory management, production planning, resource allocation.

- **Predictive Modeling:**

  -Forecasting future values based on historical data enables organizations to anticipate future trends and make proactive decisions.

  -For example, businesses can use sales forecasts to optimize inventory levels and marketing strategies.

Time series analysis and forecasting??

3.Research and Development:

-Time series analysis is crucial in scientific research and development across disciplines such as environmental science, healthcare, and engineering.

-Researchers use historical data to understand phenomena, model complex systems, and predict future outcomes.

4. Financial Analysis:

-It is used for portfolio management, risk assessment, trading strategies, and investment decision-making.

-Analysts and traders rely on historical price data to identify market trends, assess volatility, and develop trading algorithms.

Components of Time Series Data

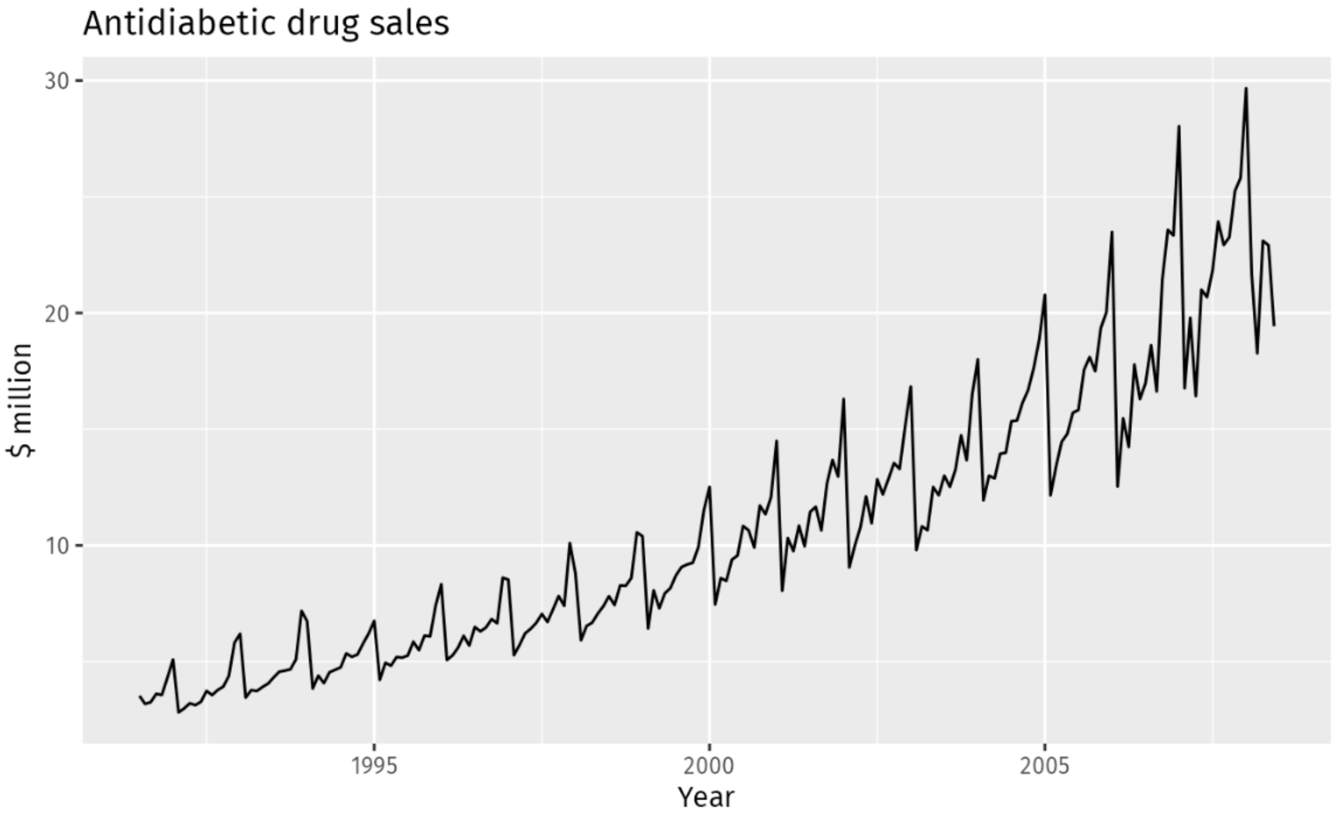
Time series can be decomposed into four components, each expressing a particular aspect of the movement of the values of the time series.

- **Secular trend** -which describe movement along the term;
- **Seasonal variations**-which represent seasonal changes;
- **Cyclical fluctuations**- which correspond to periodical but not seasonal variations
- **Irregular variations**-which are other nonrandom sources of variations of series.

Trend, seasonality, cycle and noise

Trend

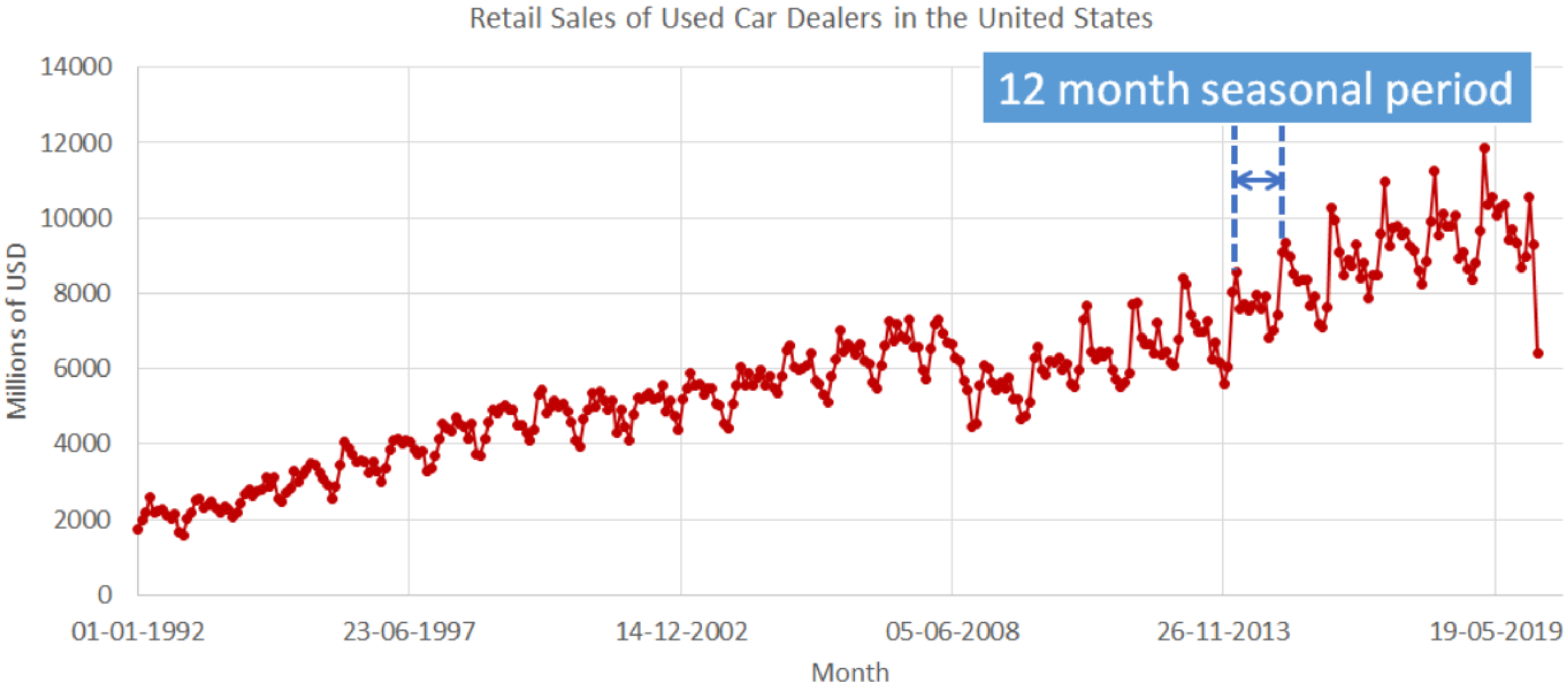
A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as “changing direction”, when it might go from an increasing trend to a decreasing trend.



Trend, seasonality, cycle and noise

Seasonal

- A seasonal pattern occurs when a time series is affected by seasonal factors such as the **time of year** or **day of week**. Seasonality is always of a fixed and known frequency.



Trend, seasonality, cycle and noise

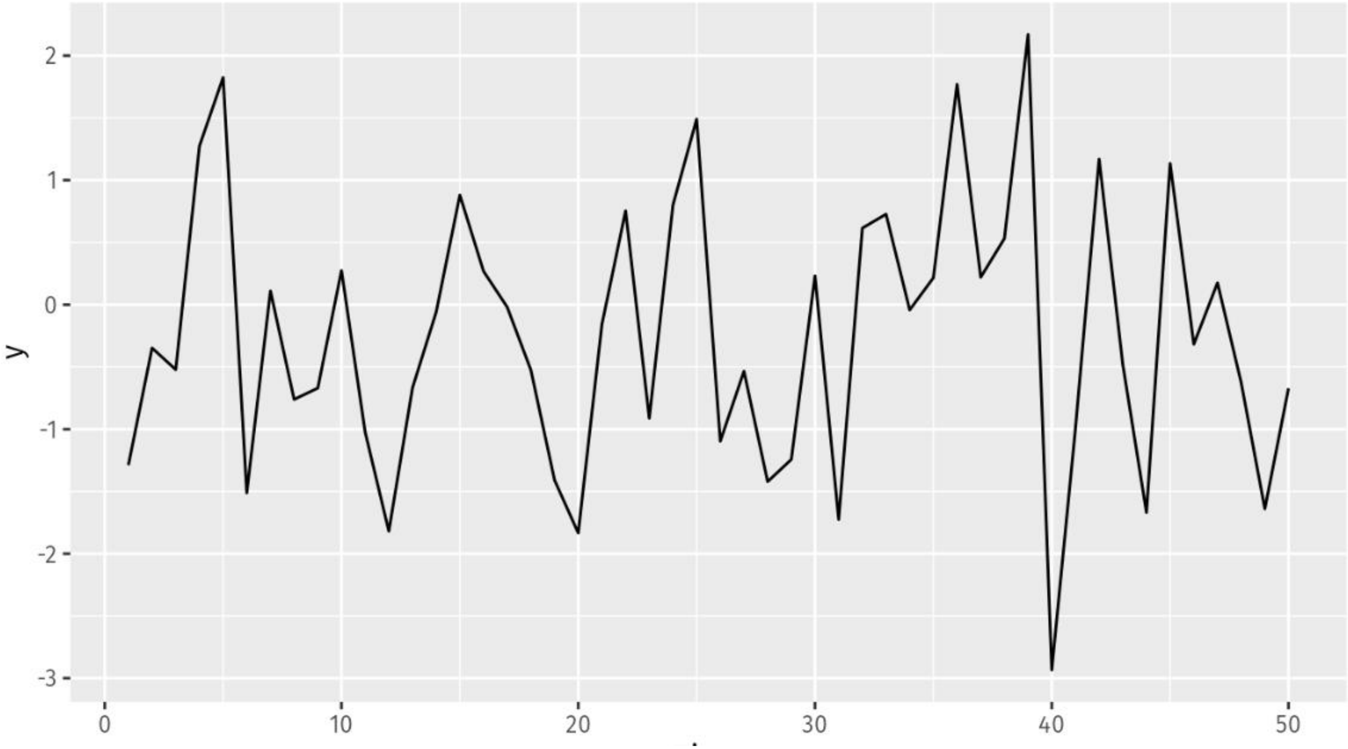
Cyclic

- A cycle occurs when the data exhibit **rises** and **falls** that are not of a fixed frequency.
- These fluctuations are usually due to economic conditions, and are often related to the “business cycle”.
- The duration of these fluctuations is usually at least 2 years.

Trend, seasonality, cycle and noise

Noise

Time series that show no autocorrelation are called **white noise**.



Forecasting

- 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 ---

  - short-term,

  -medium-term

  -long-term.

- **Short-term** – it involve predicting events only a **few time periods (days, weeks, months)** into the future.
- **Medium-term** forecasts extend from **one to two years** into the future
- **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.

Broad Types of forecasting(1)

Qualitative Forecasting Technique-

- Often subjective in nature
- Require no judgment on part of experts
- Used in situations where there is little or no historical data is available
- Example- launching a new product line

Delphi Method- Developed by RAND Corporation, 1967

- involves panel of experts
- members are physically separated to avoid dominance
- each member responds to a questionnaire, returns information
- Answers are reviewed

Broad Types of forecasting(2)

Quantitative Forecasting Technique-

- Make formal use of historical data and forecasting model
- Model summarizes patterns in the data, and expresses relationship in previous and current values of variable
- Model is used to project patterns in the data for future
- Three most used models are—

  - Regression models

  - smoothing model

  -general time series model

Terminologies

Point Estimate- 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.**

Forecast Error- Now these forecasts are almost always wrong

Prediction interval (PI)- 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.

Forecast horizon – it is the number of future periods for which forecasts must be produced. The horizon is dictated by nature of the problem. For example, in production planning, forecasts of product demand may be made on a monthly basis.

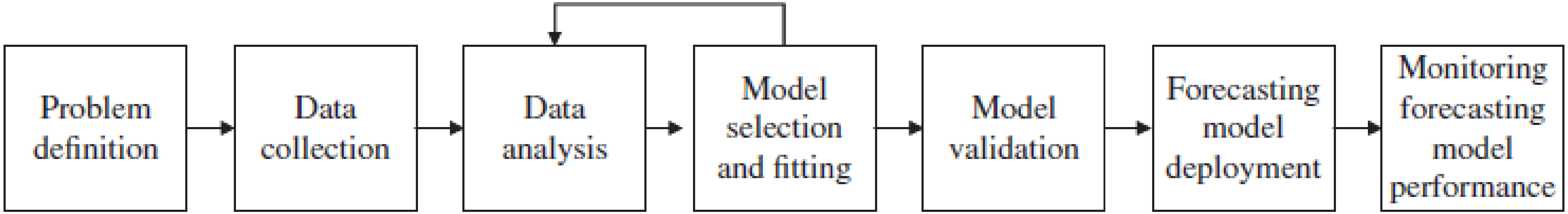
Forecast Interval- it is the frequency with which new forecasts are prepared.
For example, in production planning, forecasts are demanded on a monthly basis, for up to 3 months in the future (the lead time or horizon), and prepare a new forecast each month. Thus forecast interval is 1 month.

Forecasting Process

A process is a series of connected activities that transform one or more inputs into one or more outputs

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

Forecasting Process



Problem definition

Data collection

Data analysis

Model selection and fitting

Model validation

Forecasting model deployment

Monitoring forecasting model performance

FIGURE 1.12 The forecasting process.

Data collection and Data analysis

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

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

Data analysis is an important step to the selection of the forecasting model

- Time series plots of 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.

Model selection and fitting

- It consists of choosing one or more forecasting models and fitting the model to the data.
- **By fitting**, meaning is estimating the unknown model parameters, usually by the method of least squares.

Model Validation

- **It** 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.

Types of Time Series Data

A. Continuous Data

Continuous time series data refers to measurements or observations that can take any value within a specified range.

- **Temperature Data:** temperature recordings collected at regular intervals, such as hourly or daily measurements.
- **Stock Market Data:** data representing the prices or values of stocks, which are recorded throughout trading hours.
- **Sensor Data:** Measurements from sensors that record continuous variables like pressure, humidity, or air quality
- **Financial Data:** related to financial metrics like revenue, sales, or profit
- **Environmental Data:** collected from environmental monitoring devices, such as weather stations, wind speed, rainfall, or pollution levels.
- **Physiological Data:** physiological parameters like heart rate, blood pressure, or glucose levels recorded at regular intervals.

Continuous time series data is typically visualized using techniques such as **line plots**, **area charts**, or **smooth plots**

Types of Time Series Data

B. Discrete Data

Discrete time series data refers to measurements or observations that are limited to specific values or categories.

- **Count Data**: number of occurrences or events within a specific time. Examples-number of daily sales, number of customer inquiries per month, number of website visits per hour.
- **Categorical Data:** it falls into distinct categories or classes. This can include variables such as customer segmentation, product types, or survey responses with predefined response options.
- **Binary Data:** Data with two possible outcomes or states. For instance, a time series tracking whether a machine is functioning (1) or not (0) at each time point.
- **Rating Scales:** Data obtained from surveys or feedback forms where respondents provide ratings on a discrete scale, such as a Likert scale.

Discrete time series data is often visualized using techniques such as **bar charts**, **histograms**, or **stacked area charts**.

Time-series Data Visualization Techniques(1)

To effectively visualize time series data, the methods are:-

1.Tabular Visualization:

- presents data in a structured table format, with each row representing a specific time period and columns representing different variables.
- It provides a overview of data but may not capture trends or patterns as effectively as graphical visualizations.

2.1D Plot of Measurement Times:

- represents measurement times along a one-dimensional axis, such as a timeline.
- helps in understanding temporal distribution of data points and identifying any temporal patterns.

Time-series Data Visualization Techniques(2)

3.1D Plot of Measurement Values:

- display variation in data values over time along a single axis. Line plots and step plots are commonly used techniques for visualizing continuous time series data, while bar charts or dot plots can be used for discrete data.

4.1D Color Plot of Measurement Values:

- variation in measurement values is represented using colors on a one-dimensional axis.
- It enables the quick identification of high or low values and provides an intuitive overview of the data.

5.Bubble Plot:

- each bubble represents a data point with its size or color encoding a specific measurement value.
- allows the simultaneous representation of multiple variables and their evolution over time.

Time-series Data Visualization Techniques(3)

6.Scatter Plot:

- display relationship between two variables by plotting data points as individual dots on a Cartesian plane.
- represents one variable on the x-axis and another on the y-axis.

7.Linear Line Plot:

- It connect consecutive data points with straight lines, emphasizing **trend and continuity** of the data over time.

8.Linear Step Plot:

- it connect consecutive data points, but with vertical and horizontal lines, resulting in a **stepped appearance**.
- useful when tracking changes that occur instantaneously at specific time points.

Time-series Data Visualization Techniques(4)

9.Linear Smooth Plot:

- apply a smoothing algorithm to the data, resulting in a continuous curve that captures the overall trend while reducing noise or fluctuations.
- It helps in **visualizing long-term patterns** more clearly.

10.Area Chart:

- fill area between the line representing the data and the x-axis, emphasizing the cumulative value or distribution over time. They are commonly used to visualize stacked time series data

11.Horizon Chart:

- condense time series data into a compact, horizontally layered representation.
- compares multiple time series data on a single chart, optimizing screen space usage.

Time-series Data Visualization Techniques(5)

12.Bar Chart:

- represent discrete time series data using **rectangular bars**, with the height of each bar indicating the value of a specific measurement.
- compares values between different time periods or categories.

13.Histogram:

- Display distribution of continuous or discrete time series data by **dividing range of values into equal intervals (bins)**
- Representing frequency or count of data points falling within each bin.

NUMERICAL DESCRIPTION OF TIME SERIES DATA

Stationary Time Series-

A time series is said to be **strictly stationary** if its properties are not affected by a change in the time origin.

- If the joint probability distribution of the observations *yt*, *yt*+1,…, *yt*+*n* is exactly the same as the joint
- probability distribution of the observations *yt*+*k*, *yt*+*k*+1,…, *yt*+*k*+*n* then the time series is strictly stationary.
- When *n* = 0 the stationarity assumption means that the probability distribution of *yt* is the same for all time periods
- Stationary implies a type of statistical **equilibrium** or **stability** in the data.

Mean,

$$\mu_y = E(y) = \int_{-\infty}^{\infty} y f(y) dy$$

Variance

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

Example- Stationary Time Series

The pharmaceutical product sales and chemical viscosity readings time series data

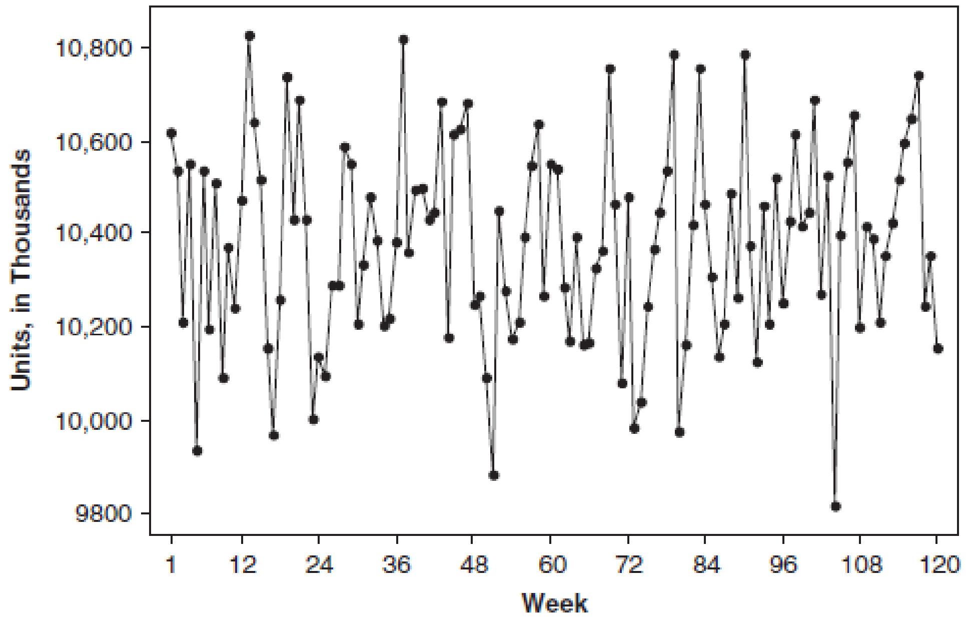


FIGURE 1.2 Pharmaceutical product sales.

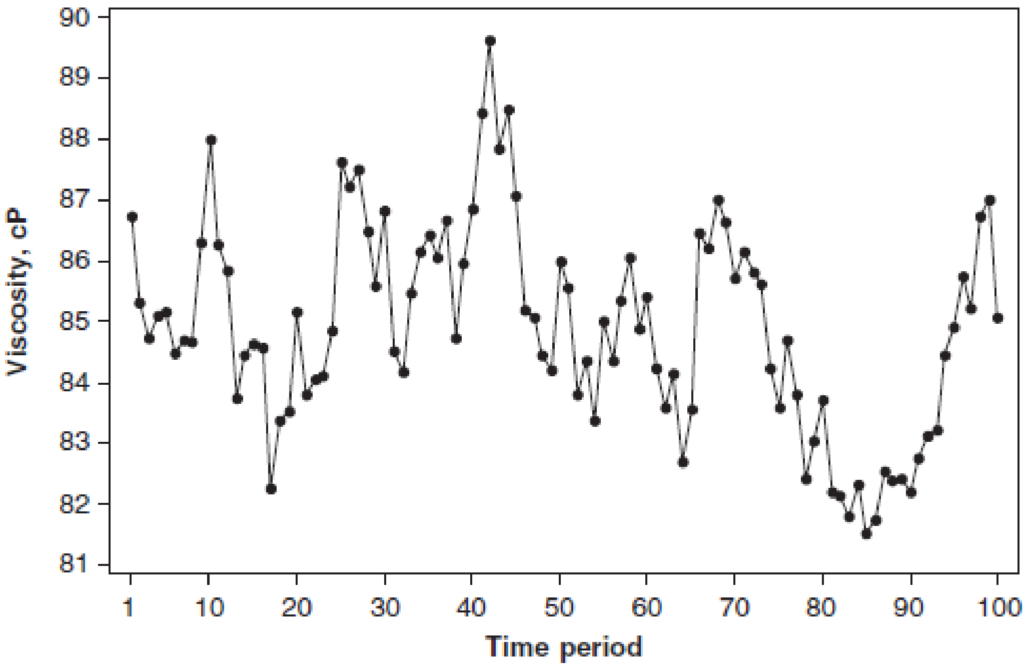


FIGURE 1.3 Chemical process viscosity readings.

NUMERICAL DESCRIPTION OF TIME SERIES DATA

Autocovariance Functions

- The covariance between *yt* and its value at another time period, say, *yt+k* is called the **autocovariance** at lag *k*, defined by

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

- The collection of the values of *γk*, *k* = 0, 1, 2,… is called the **autocovariance function**.

Autocorrelation Functions (ACF)

autocorrelation coefficient at lag *k* for a stationary time series is-

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

Example- Positively correlated and Unrelated data

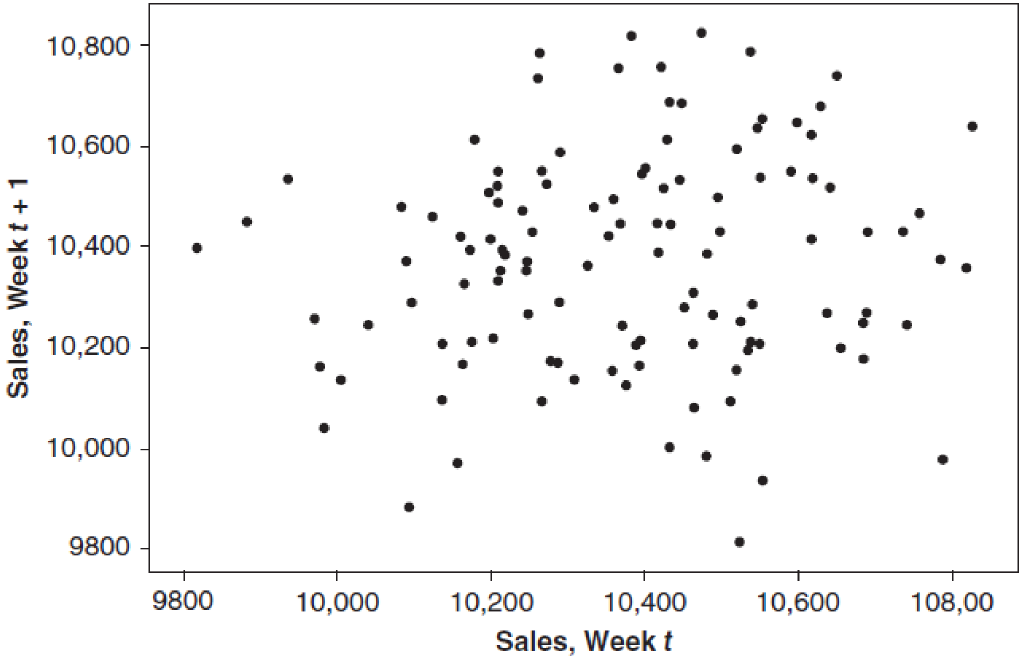


FIGURE 2.10 Scatter diagram of pharmaceutical product sales at lag $k = 1$.

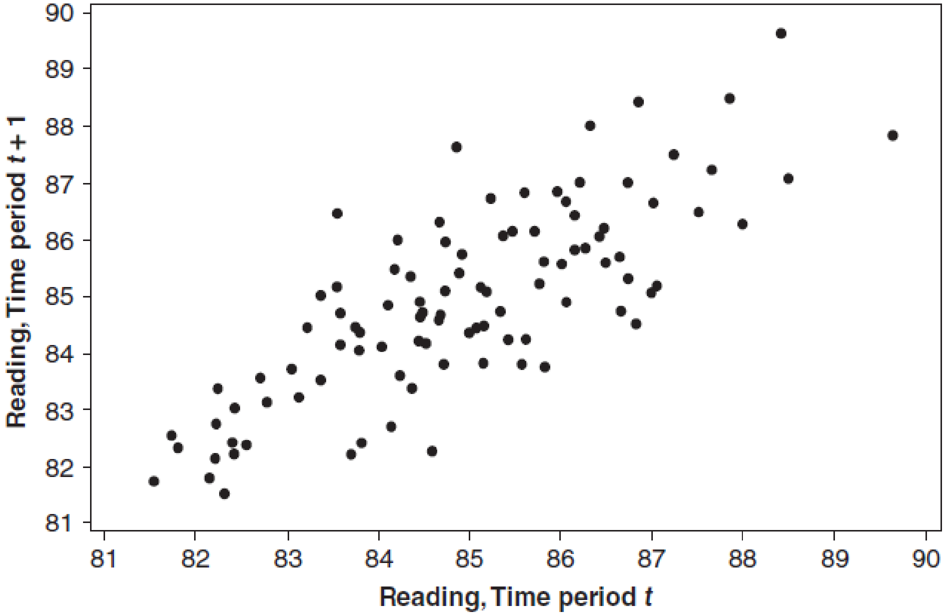


FIGURE 2.11 Scatter diagram of chemical viscosity readings at lag $k = 1$.

Review questions

1. Why is forecasting an essential part of the operation of any organization or business?
2. What is a time series? Explain the meaning of trend effects, seasonal variations, and random error.
3. Explain the difference between a point forecast and an interval forecast.
4. What do we mean by a causal forecasting technique?