MINING TIME SERIES DATA

STQD6414 PERLOMBONGAN DATA

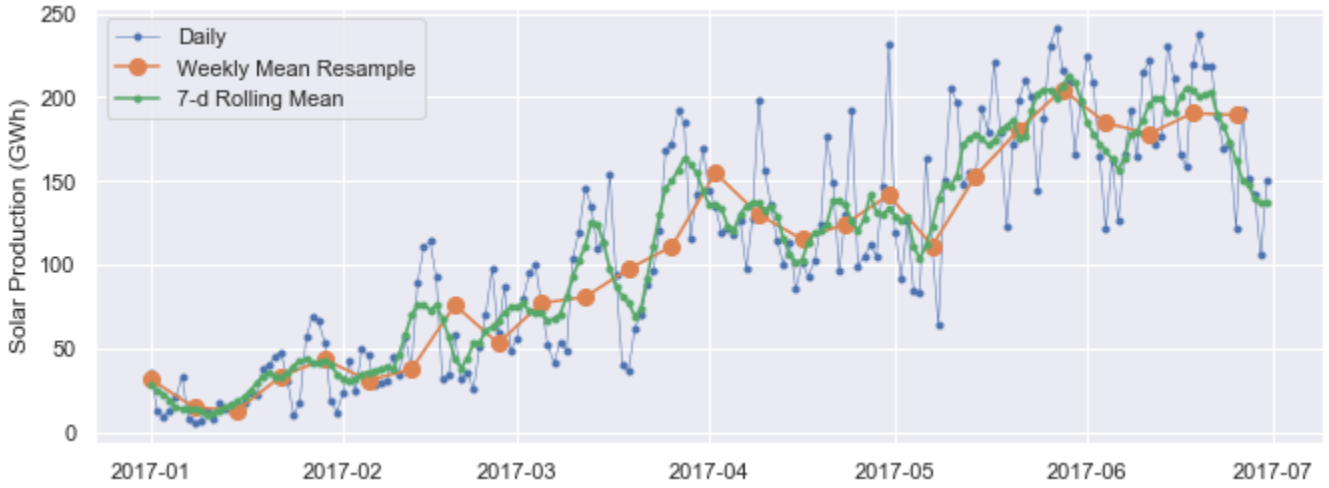
Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Time series data is a collection of observations obtained through repeated measurements over time.
- A time series data tracks a sample over some particular periods.
- Time series data is different from cross-sectional data, which collects a number of different samples at the same time.
- In general, time series data is useful for tracking the characteristics and behaviors of daily, hourly, weekly, or yearly related to some particular events.



INTRODUCTION:

- Example of time series data:

i) Monthly retail sales

ii) Movement of Stock Price

iii) Weather Forecasting

iv) Economic indicators over time.

v) And many more.

- Time Series Mining includes the following topics (in time domain):

i) Time Series Decomposition:

- Time Series data can be decomposed into the components of trend, seasonal, cyclical and random term.

INTRODUCTION:

ii) Time Series Forecasting:

- Build a statistical model and then use it to predict future values.
- There are many popular time series models such as time series regression models, ARIMA models, GARCH models, non-linear models and etc.

iii) Time Series Clustering:

- Time series clustering is the process to segmenting multiple time series data into several clusters based on the nature of similarity or distance.
- Common clustering techniques are discussed in the Machine Learning course.

iv) Time Series Classification:

- Time series classification aims to construct a classification model based on labeled time series data.
- Common data classification techniques are discussed in the Machine Learning course.

TIME SERIES DATA IN R:

- There are several classes for time series data in R. Among them:

i) Date class

ii) ts class

iii) POSIX class

iv) timeSeries class

v) zoo class

vi) xts class

vii) And many more.

- The time series data entry in R is usually in the form of daily, weekly, monthly, annual or quarterly data.
- Example: Frequency=12 and start=c(2011,3) states that the data is a monthly time series starting from March 2011.

CREATING DATE AND TIME OBJECTS:

- R adopted the ISO 8601 format for time.
- Commonly, as.Date() function is used the create date in R.
- Date object in R refer to character string.
- The character string has to obey a format that can be defined using a set of symbols
- Example: 12 January, 2022

i) %Y: 4-digit year (2022)

ii) %y: 2-digit year (22)

iii) %m: 2-digit month (01)

iv) %d: 2-digit day of the month (12)

v) %A: weekday (Wednesday)

vi) %a: abbreviated weekday (Wed)

vii) %B: month (January)

viii) %b: abbreviated month (Jan)

TIME SERIES DECOMPOSITION:

- Time Series Decomposition is a method to decompose time series data into a structural component which are:

i) Trend,

ii) Seasonality,

iii) Cyclical.

- And also non- structural component:

i) Irregular (random) components.

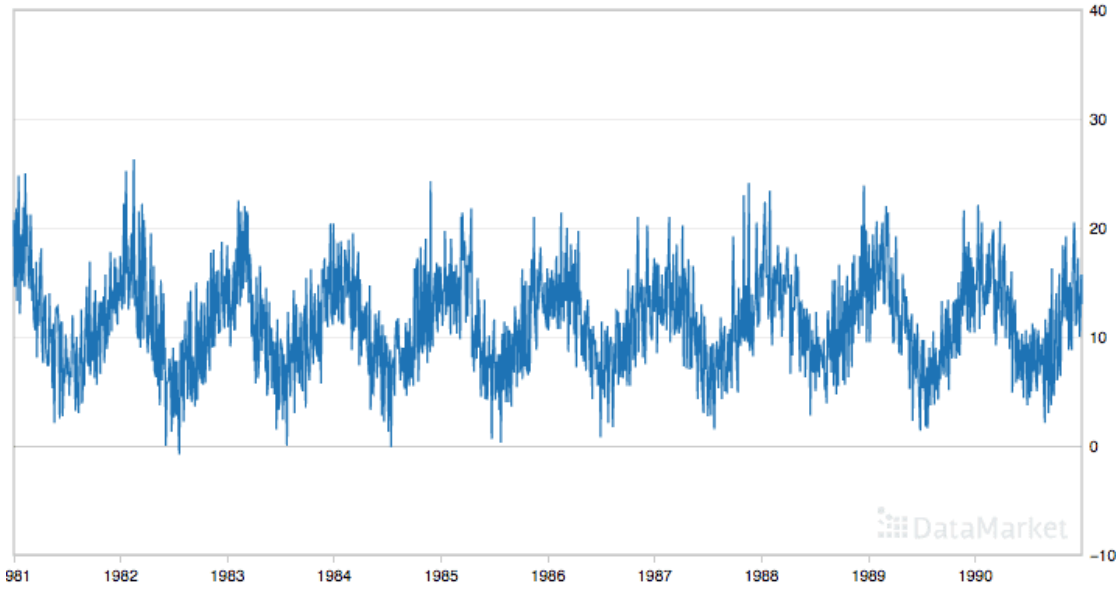
- Decomposition provides information about the behaviors of time series data for better understanding during time series analysis and forecasting.

THE TREND COMPONENT:

- The trend component refers to increasing or decreasing trend in a data.
- Depending on the series characteristics, a trend could have either linear, polynomial or exponential growth.

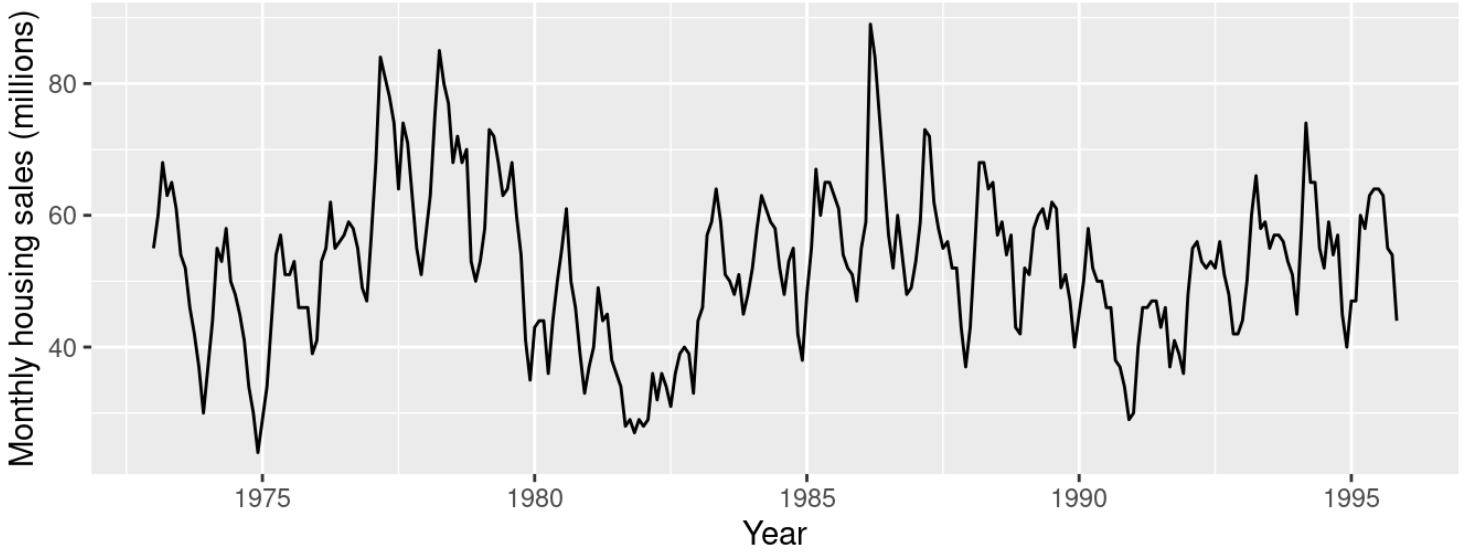
THE SEASONAL COMPONENT:

- The seasonal component is a seasonal variation that occurs periodically.
- Example:
- Hourly seasonality, which is derived from parameters such as; sunlight hours and temperatures throughout the day.
- Weekly seasonality, which depends on the day of the week.
- Monthly seasonality, which is related to the season of the year.



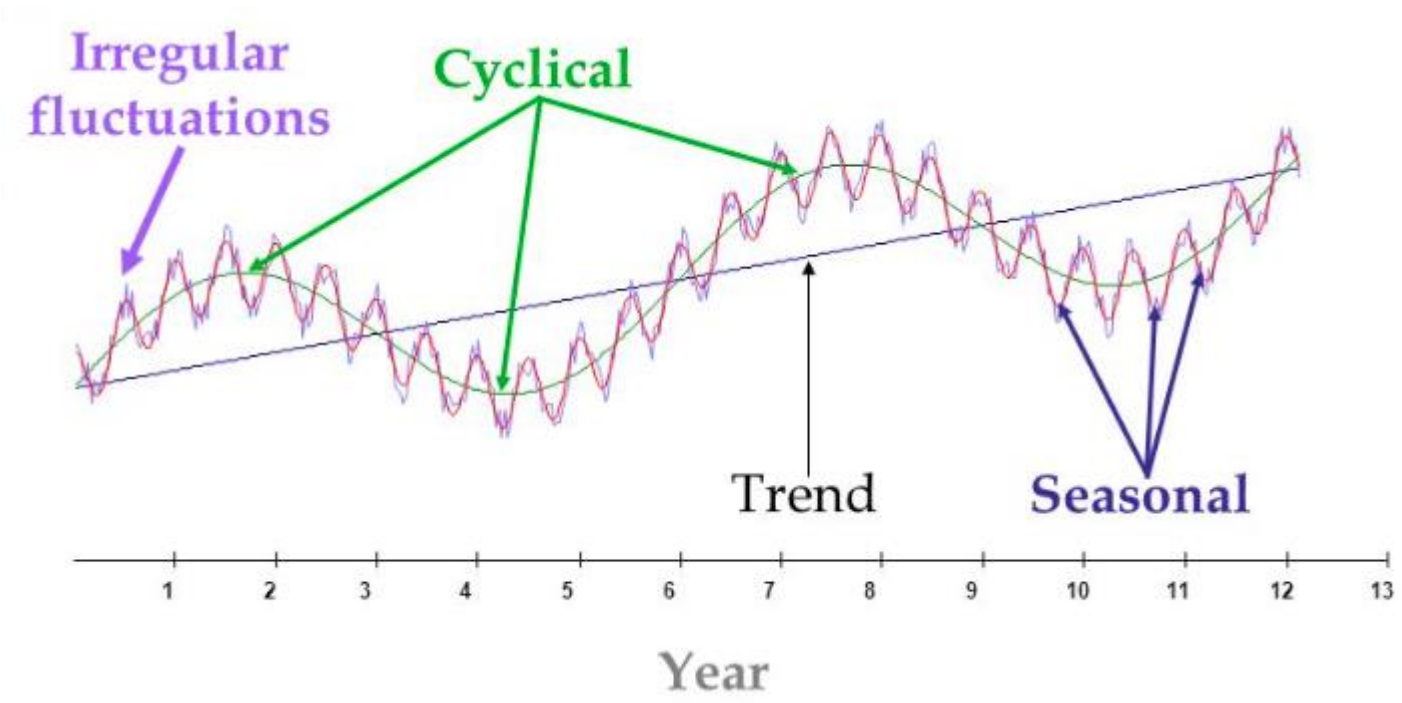
THE CYCLE COMPONENT:

- A cycle is a sequence of repeatable events over a long period of time.
- Unlike the seasonal pattern, cycles do not necessarily occur at equally spaced time intervals, and their length could change from cycle to cycle.



THE IRREGULAR COMPONENT:

- Irregular component is a remainder between the series and structural components (trend, seasonality, and cyclical).
- Its provides an indication of irregular events in the series (random components).
- Its show non-systematic patterns or events in the data.

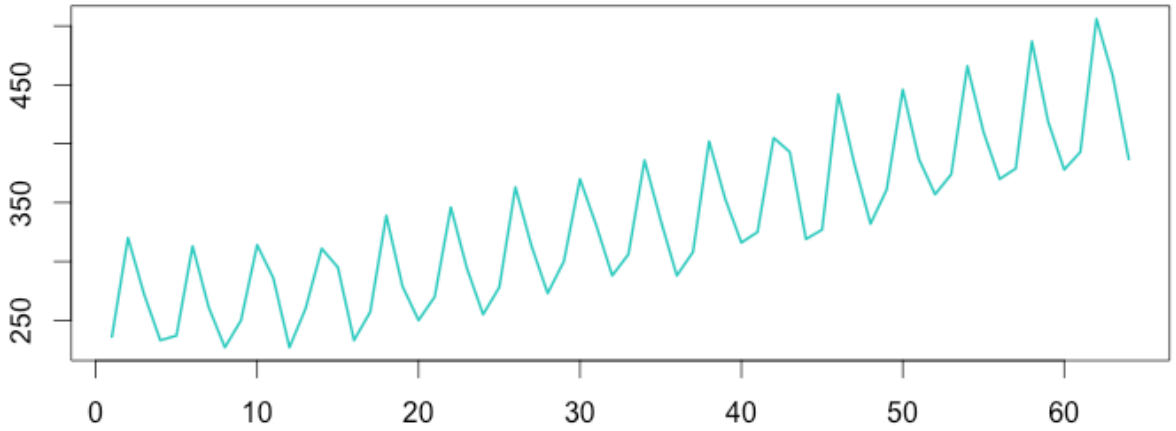


TYPES OF DECOMPOSITION:

- There are two types of time series decomposition, which are:

i) Additive Decomposition:

- Additive structure of time series exists when there is a growth in the trend, or if the amplitude of the seasonal component roughly remains the same over time.



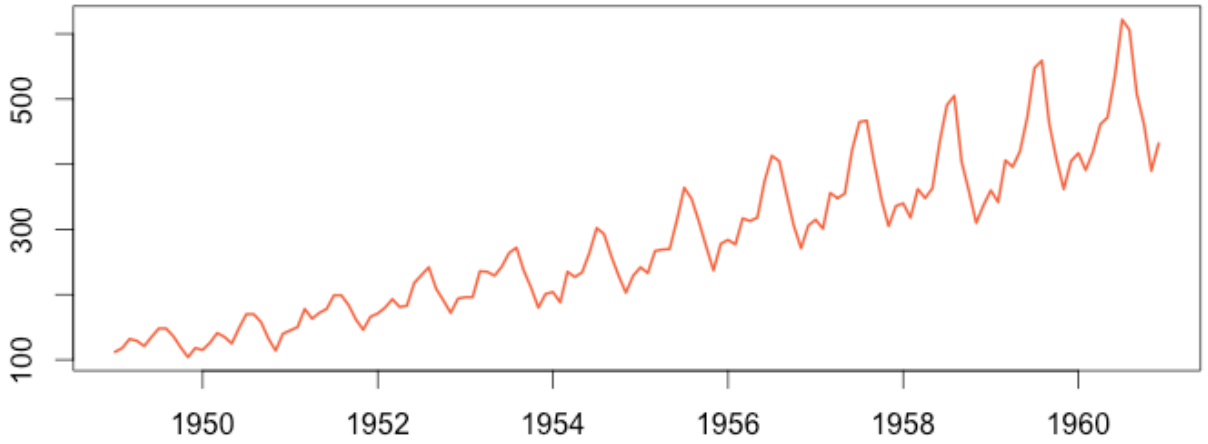
- An additive structure in time series can be represents as the following equation:

$$Y_t = T_t + S_t + C_t + I_t$$

TYPES OF DECOMPOSITION:

ii) Multiplicative Decomposition:

- Multiplicative structure of time series exists whenever the growth of the trend or the magnitude of the seasonal component increases or decreases by some multiplicity from period to period over time.



- Multiplicative structure of time series:

$$Y_t = T_t \times S_t \times C_t \times I_t$$

- Multiplicative structure in time series can be converts to additive structure using Box-Cox Transformation.

TIME SERIES FORECASTING:

- Time series model is used to predict future events based on past data.

- Example: We want to predicts the opening price of a stock based on past stock performance.

- Two basic models (time domain) for time series forecasting are the autoregressive moving average (ARMA) and the autoregressive integrated moving average (ARIMA).

- The fitted ARIMA model to univariate time series data can be used for forecasting.

BASIC ARIMA MODEL:

- The ARIMA model is formed from a combination of Autoregressive (AR) and Moving Average (MA) models.

(1) Autoregresive model with *p*-order, AR(*p*):

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

- y_t depends on the past *p* values of observations.

(2) Moving Average Model with *q-order,* MA(*q*):

$$y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

- y_t depends on the past *q* residual values.

BASIC ARIMA MODEL:

(3) The combination of the AR(p) and MA(q) models yields the ARMA(p, q) model:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

- y_t depends on the past p values of observations and also their past q residual values.
- If the data is not stationary, a differentiation technique is performed on the data to make it stationary.
- i-order of differencing create ARIMA(p, i, q) model.

STATIONARITY OF TIME SERIES:

- "Stationarity" is a fundamental property of most time series statistical models.

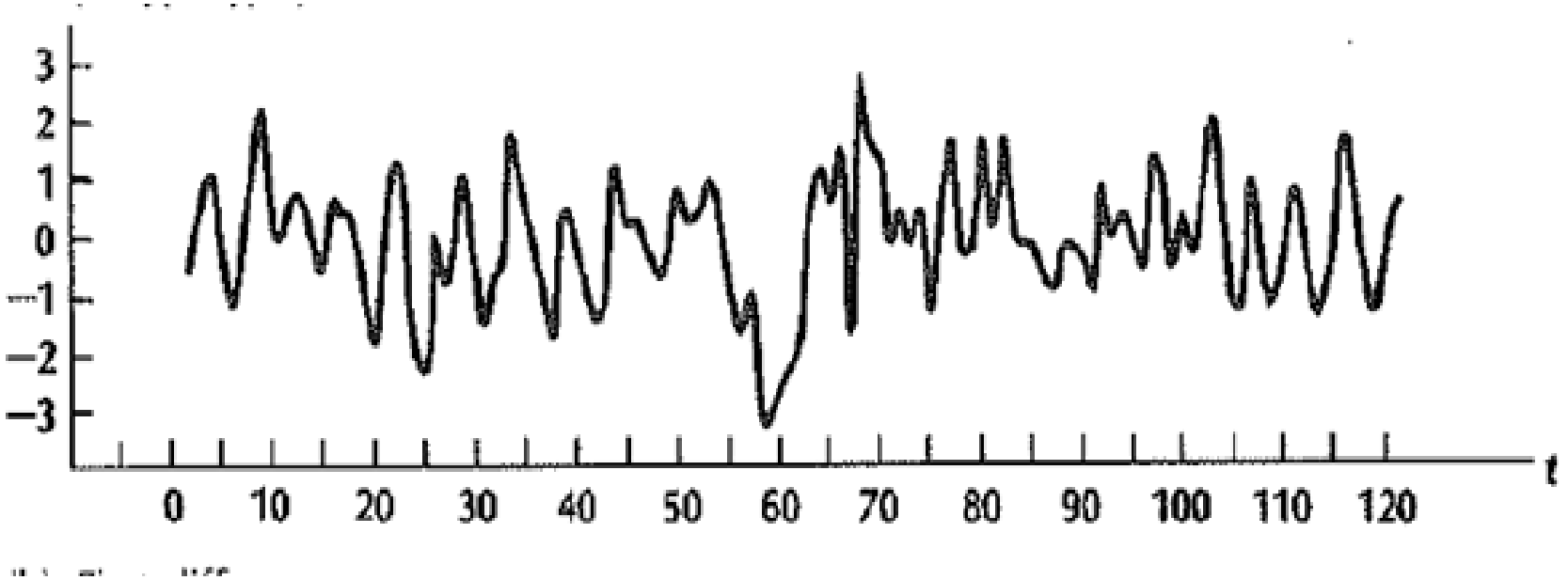
- A time series y_t is said to be stationary if it satisfies the following conditions:

$$(1)\ \ E(y_t)=u_y \ \ for\ all\ t.$$

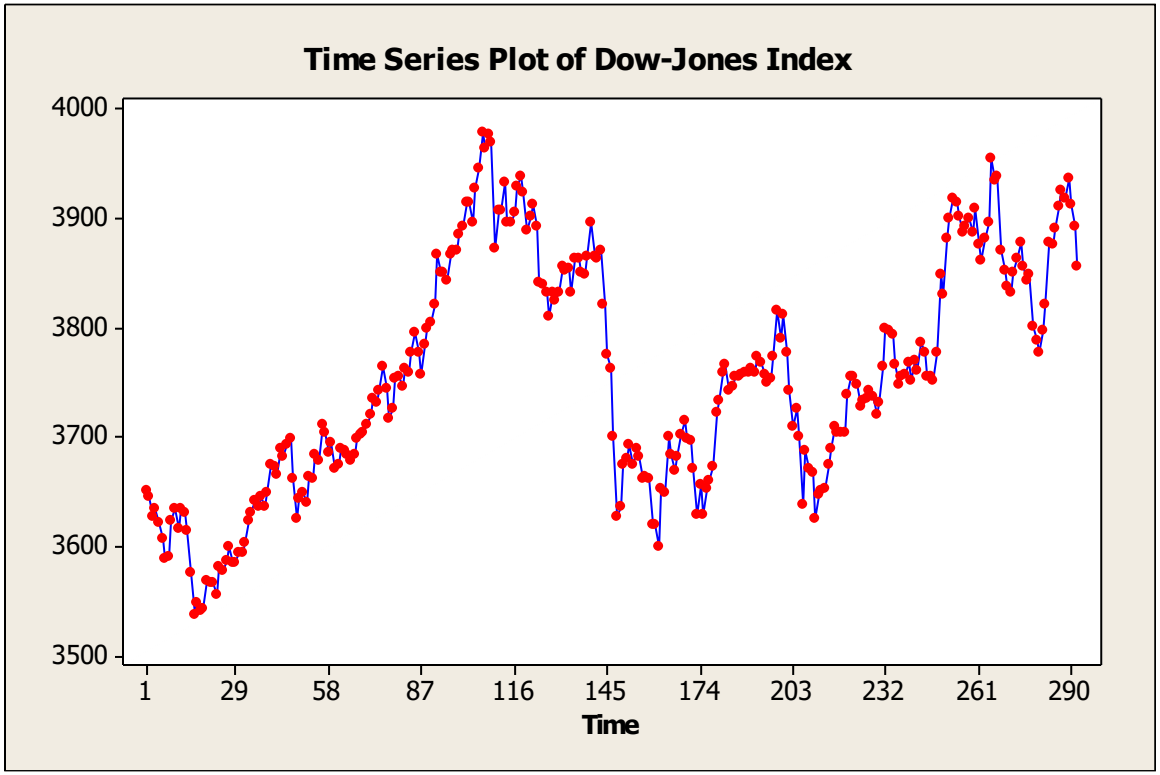
$$(2)\ Var(y_t)=E[(y_t-u_y)^2]=\sigma_y^2 \ \ for\ all\ t.$$

$$(3)\ Cov(y_t, y_{t-k})=\gamma_k \ \ for\ all\ t.$$

EXAMPLE: STATIONARY TIME SERIES:



EXAMPLE: NON-STATIONARY SERIES:



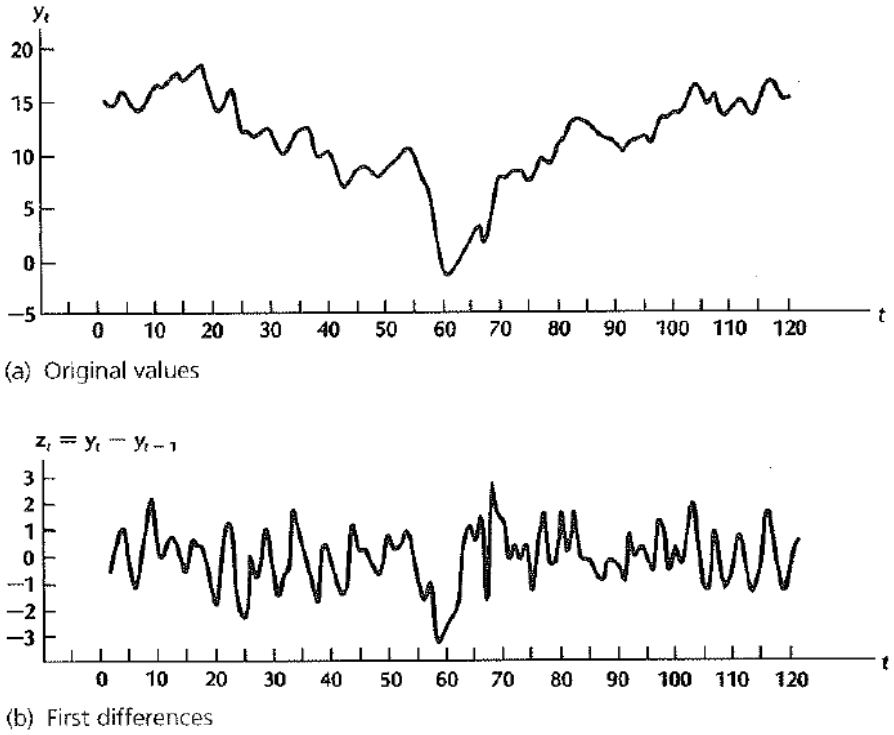
DIFFERENCING:

- Non-stationary series can be transform to stationary series through differencing.

- Example:

y_t is not stationary, but

$z_t = y_t - y_{t-1}$ is stationary



(a) Original values

(b) First differences

DIFFERENCING:

- Differencing continues until stationarity is achieved.

$$\Delta y_t = y_t - y_{t-1}$$

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$

- The number of times that the original series must be differenced in order to achieve stationarity is called the order of integration, denoted by *i*.

- In practice, it is not necessary to go beyond second difference, because real data generally involve only first or second level non-stationarity.

ARIMA MODEL IDENTIFICATION:

- Once the data is stationary, we can proceed to identify ARIMA model through a visual inspection of the sample autocorrelation (AC) and partial sample autocorrelation (PAC) functions.
- For the series $y_1, y_2, \ldots, y_n$, the sample autocorrelation at lag *k* is

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

- Where $\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n}$.

- The sample partial autocorrelation at lag *k* is:

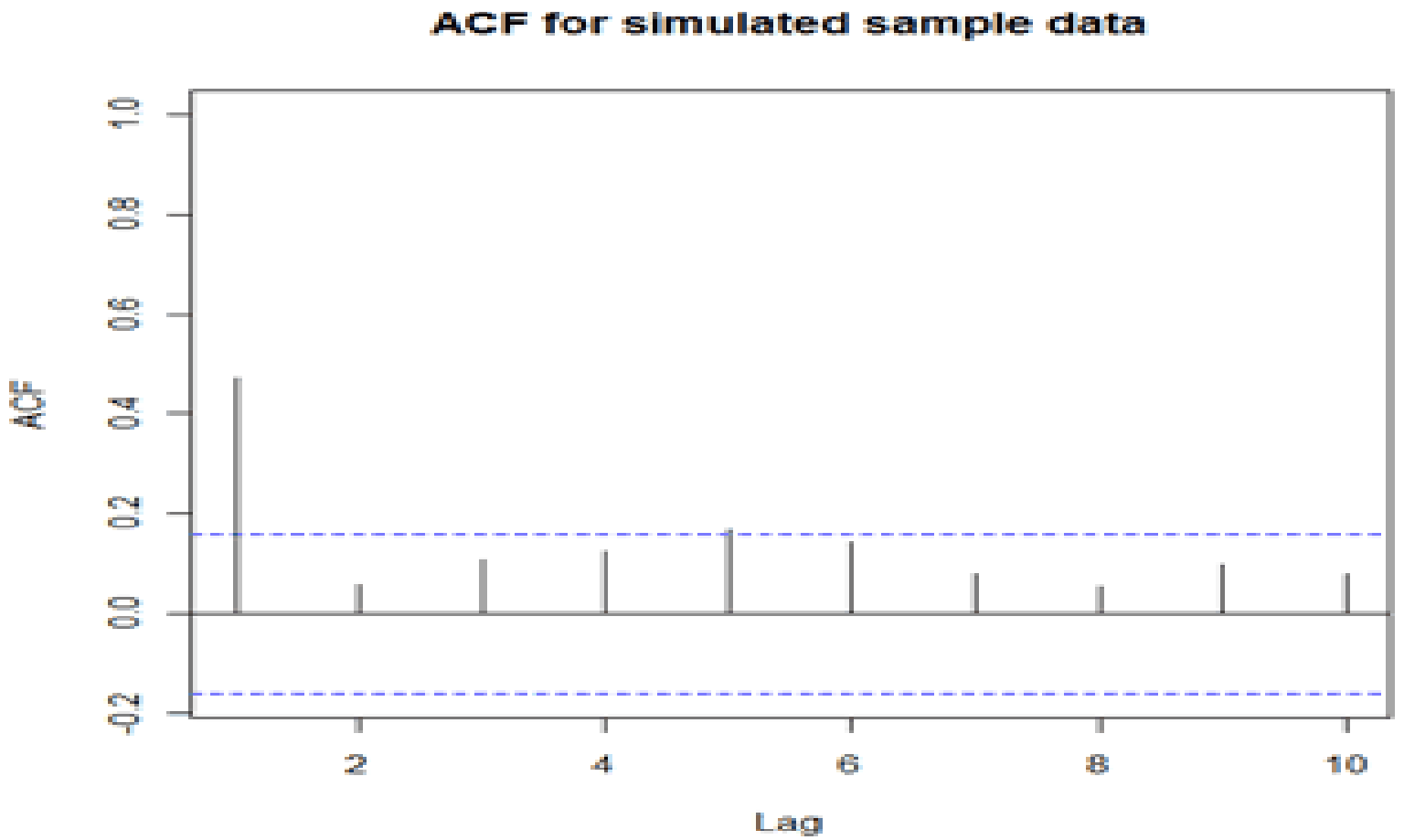
$$r_{kk} = \begin{cases} r_1 & \text{if } k = 1, \\ \dfrac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} \cdot r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} \cdot r_k} & \text{if } k = 2,3,\ldots \end{cases}$$

- Where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ for j = 1, 2, …, k-1.

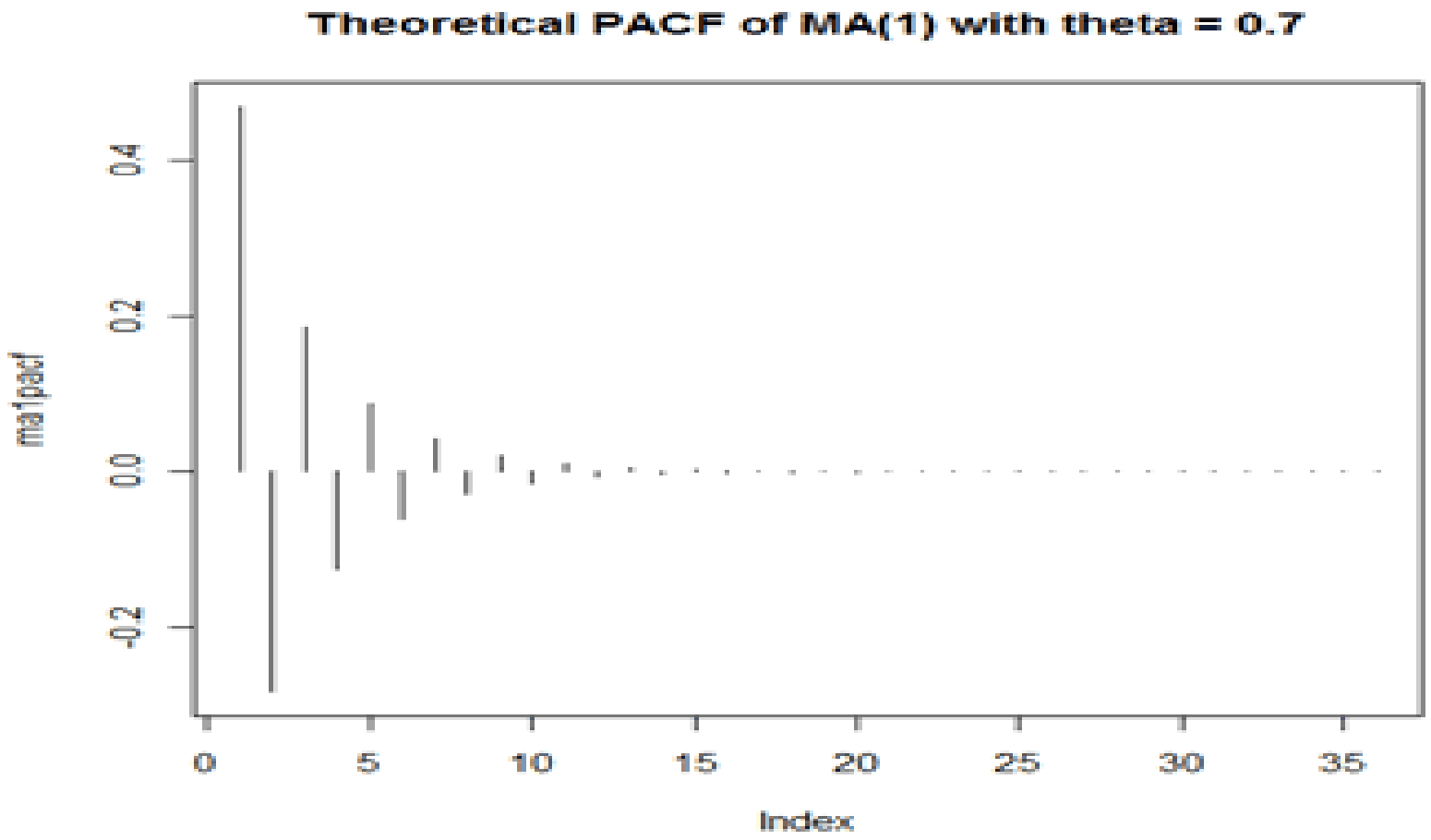
THE BEHAVIORS OF ACF AND PACF:

Model	AC	PAC
Autoregressive of order p $y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$	Dies down	Cuts off after lag p
Moving Average of order q $y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$	Cuts off after lag q	Dies down
Mixed Autoregressive-Moving Average of order (p,q) $y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$	Dies down	Dies down

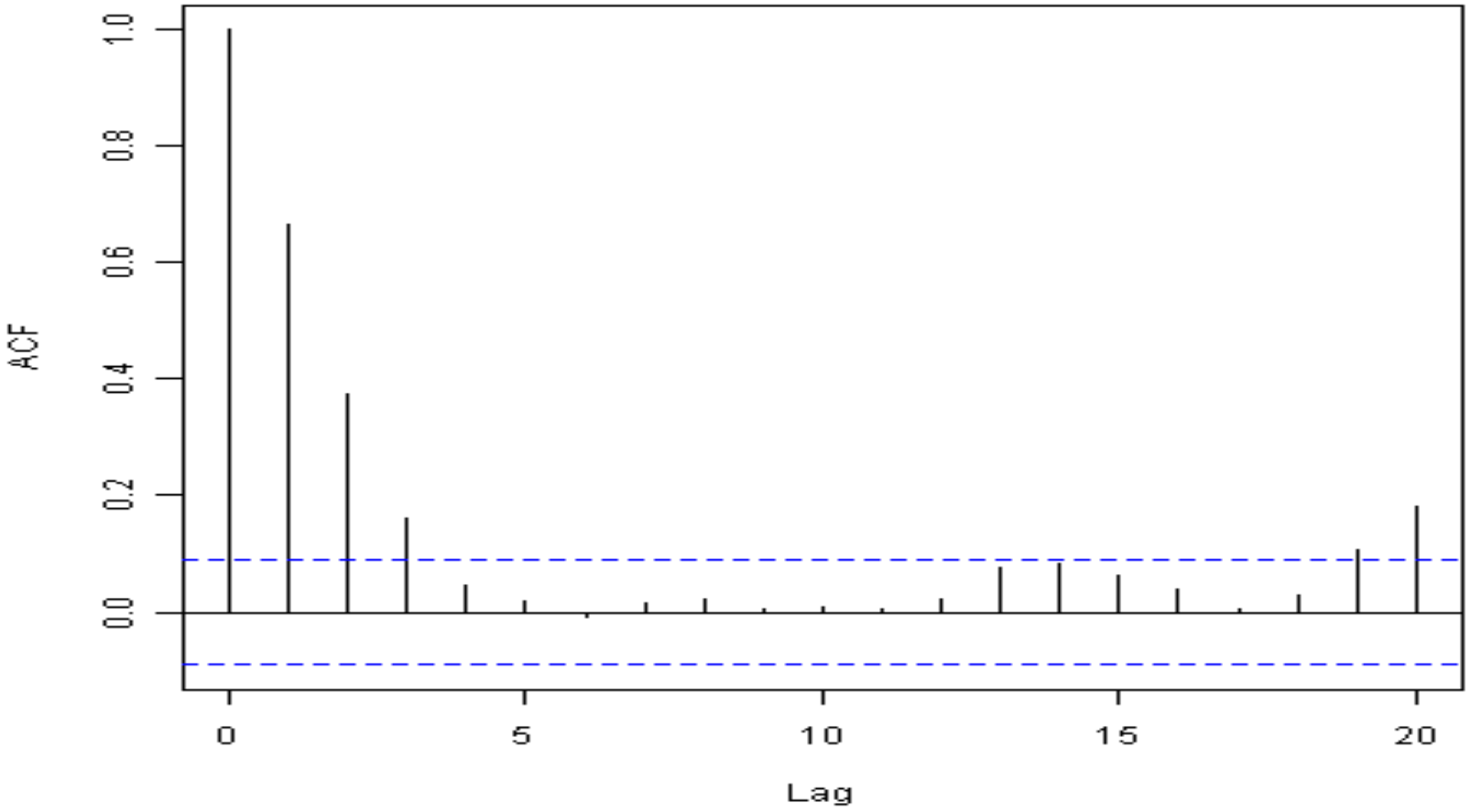
EXAMPLE: 1ST ORDER MOVING AVERAGE MODEL, MA(1)



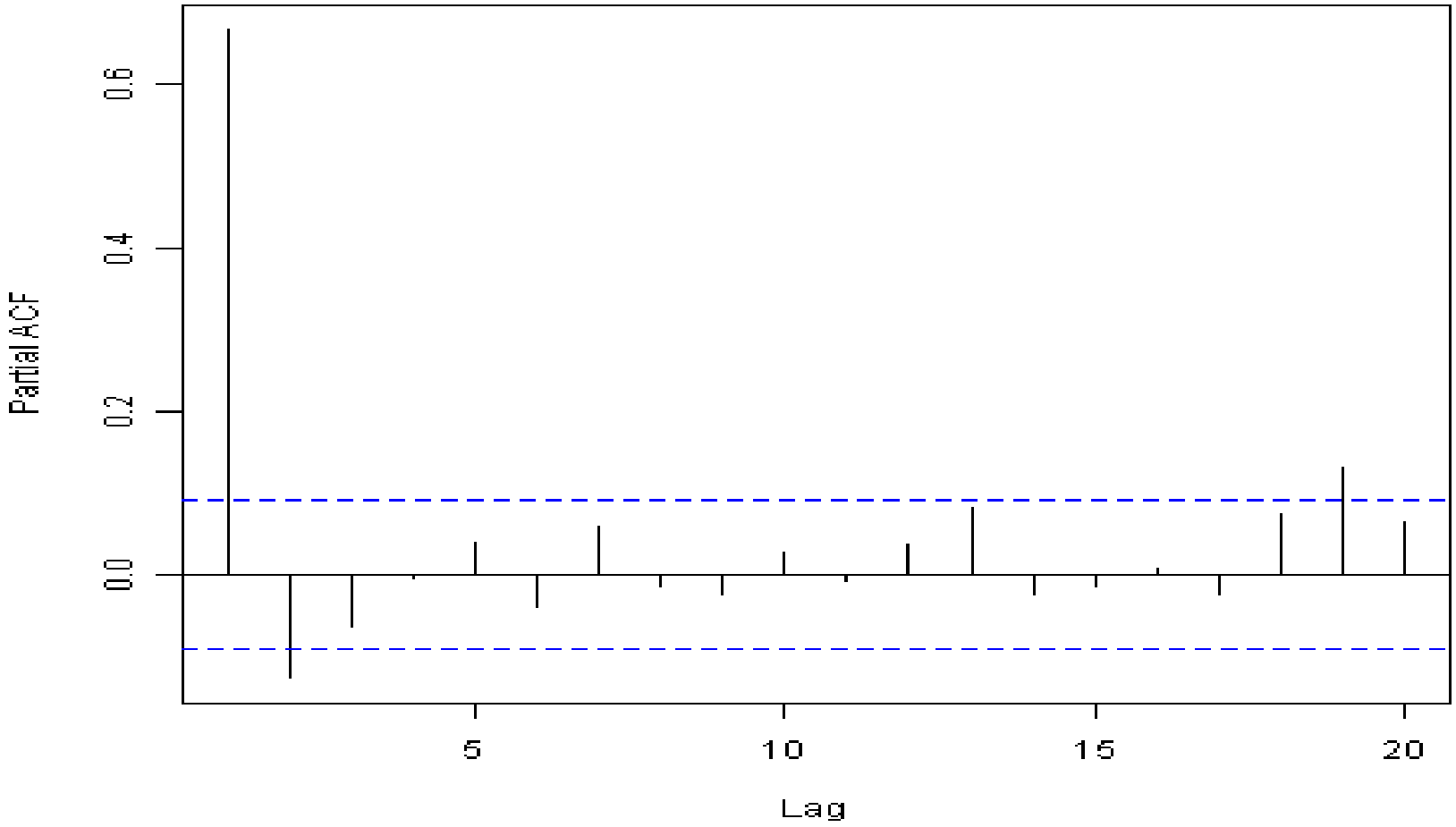
EXAMPLE: 1ST ORDER MOVING AVERAGE MODEL, MA(1)



EXAMPLE: 2ND ORDER AUTOREGRESSIVE MODEL, AR(2)



EXAMPLE: 2ND ORDER AUTOREGRESSIVE MODEL, AR(2)



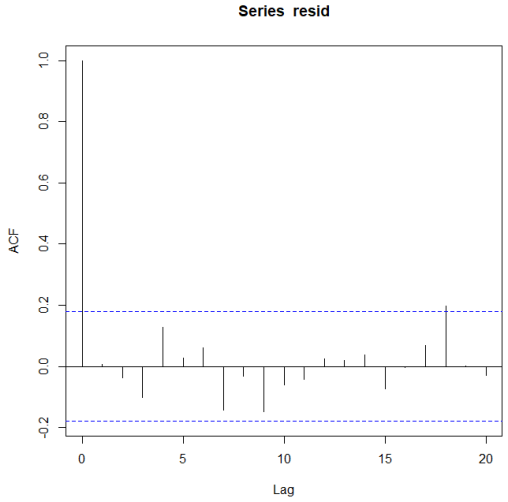
DIAGNOSTIK MODEL:

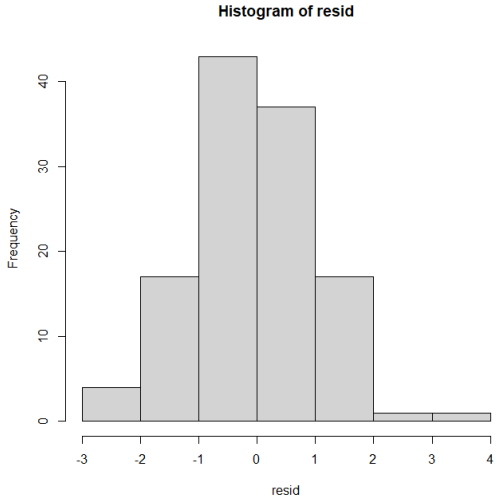
- Once, ARIMA model that has been fitted to the data, it needs to checked their mathematical assumptions behind the ARIMA model.
- This can be done through a residual analysis.
- Residual analysis:

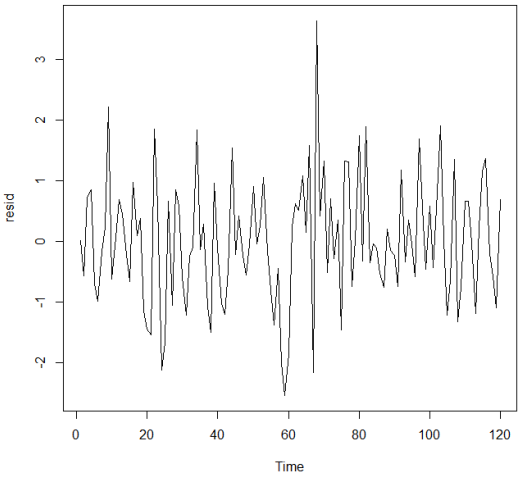
i) Residual should not be is uncorrelated.

ii) Residual is distributed approximately normally.

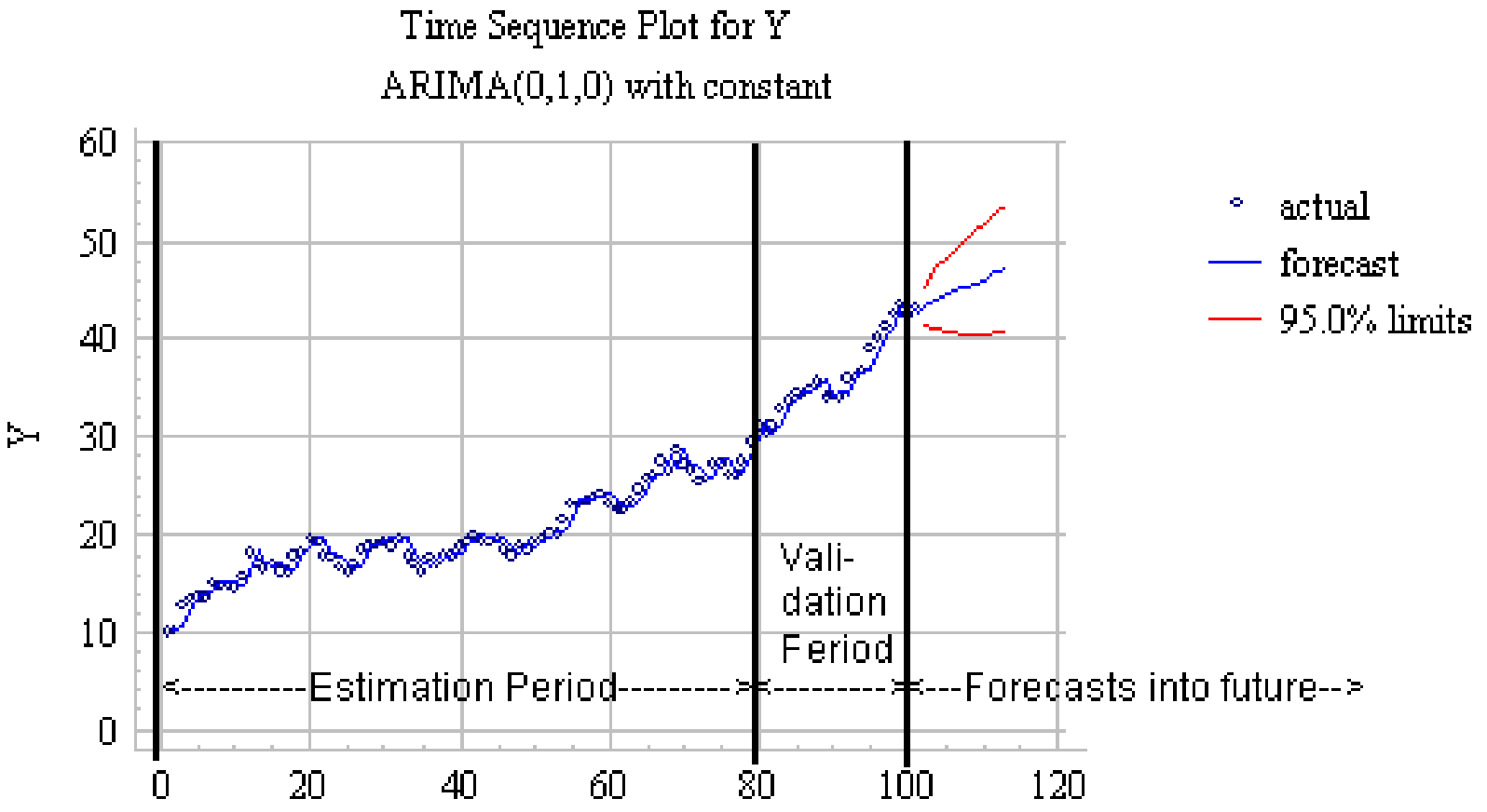
iii) The variance of the residual is constant over time.







EXAMPLE: MODEL FITTED & FORECASTING



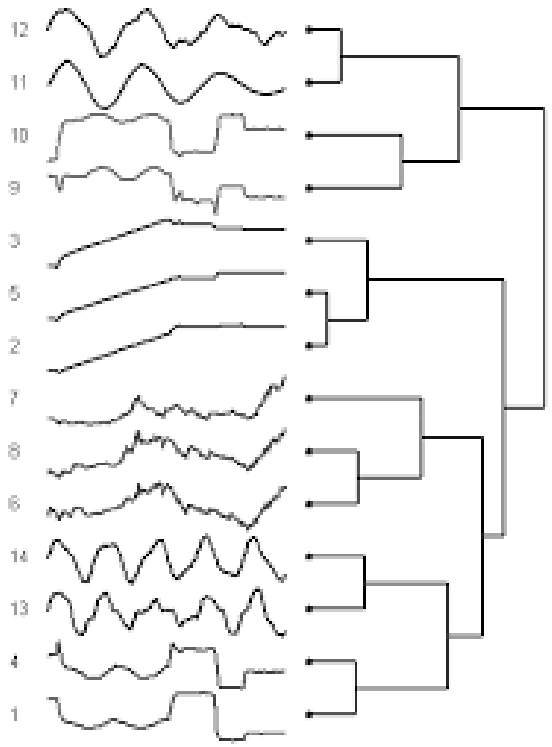
SUMMARIZATION OF ARIMA MODELING:

1) Plot the time series data and determine whether the data is having a stationary properties or not.
2) Determine the appropriate ARIMA model based on the ACF and PACF plots or AIC measure.
3) Fit the ARIMA model to the data.
4) Perform residual analysis for model validation.
5) Use the fitted ARIMA model to forecast future values.
6) Compute confidence interval for forecast values.

TIME SERIES CLUSTERING:

- Time series clustering is the process of segmenting multiple time series data into several clusters based on the nature of similarity or distance.
- Time series in the same cluster will have high similarity characteristics.
- Time series in different clusters will have low similarity characteristics.

- Among the cluster features of the time series:
  - (i) Normal
  - (ii) Cyclic
  - (iii) Increasing trend
  - (iv) Decreasing trend
  - (v) Upward shift
  - (vi) Downward shift



MEASUREMENT OF DIFFERENCE :

- There are four main approach to measure the distance in time series clustering, such as:

i) Model-free approaches

- Euclidean distance.
- Minkowski distance.
- Manhattan distance.
- Dynamic Time Wrapping (DTW) distance.
- Correlation-based distances.
- Autocorrelation-based distances.
- Periodogram-based distances.
- And many more.

MEASUREMENT OF DIFFERENCE:

ii) Model-based approaches

- Piccolo distance.
- Maharaj distance.
- Central-based distance.

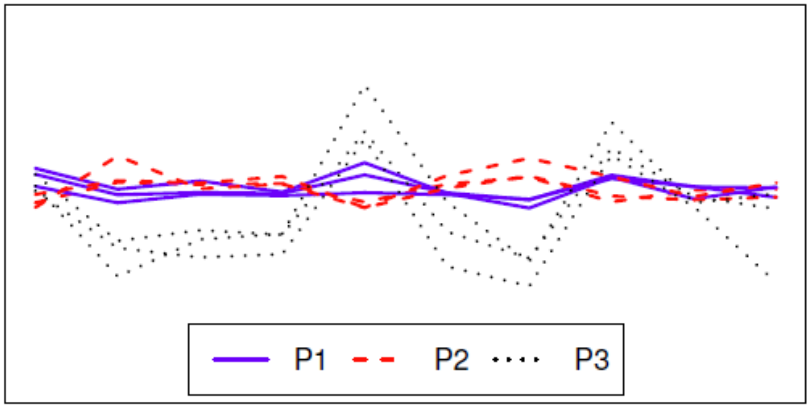
iii) Complexity-based approaches:

- Compression-based dissimilarity measures.
- Permutation distribution clustering.
- A complexity-invariant dissimilarity measure.

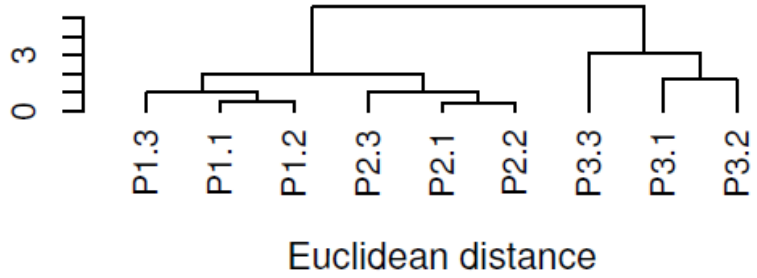
iv) Prediction-based approaches.

EXAMPLE:

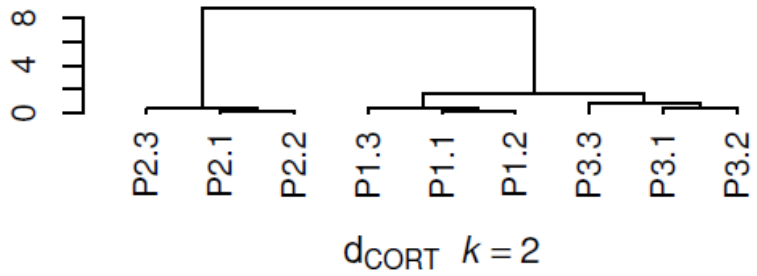
(a)


P1
P2
P3

(b)


0
3
P1.3
P1.1
P1.2
P2.3
P2.1
P2.2
P3.3
P3.1
P3.2
Euclidean distance

(c)


0
4
8
P2.3
P2.1
P2.2
P1.3
P1.1
P1.2
P3.3
P3.1
P3.2
d_{CORT} $k = 2$

TIME SERIES CLASSIFICATION:

- Time series classification aims to construct a classification model based on labeled time series.
- Then, the determined classification model will be used to predict new unlabeled time series data.
- New features extracted from the time series can help to improve the performance of the classification model.

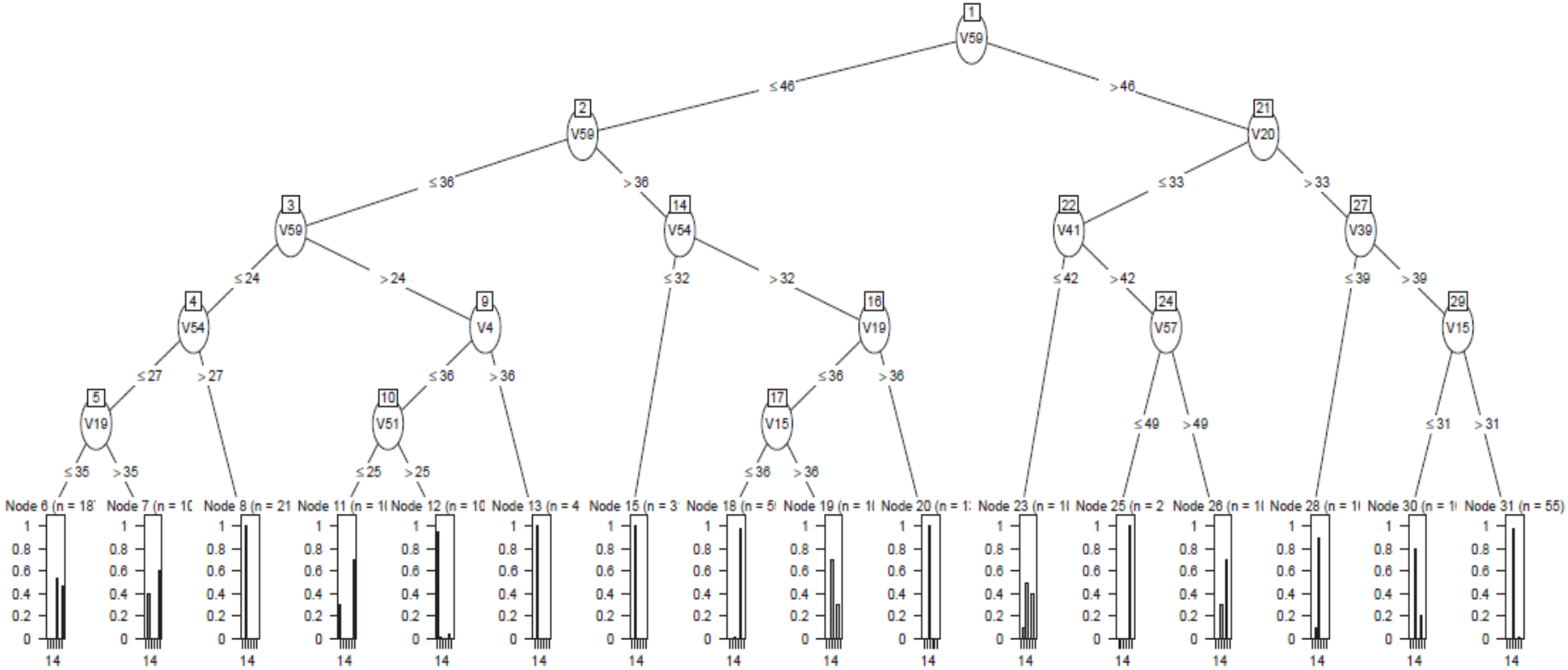
- Among the techniques used for feature extraction in Time Series are:

i) Singular Value Decomposition (SVD)

ii) Discrete Fourier Transform (DFT)

iii) Discrete Wavelet Transform (DWT)

EXAMPLE:


1
V59
≤ 46
> 46
2
V59
21
V20
≤ 36
> 36
≤ 33
> 33
3
V59
14
V54
22
V41
27
V39
≤ 24
> 24
≤ 32
> 32
≤ 42
> 42
≤ 39
> 39
4
V54
9
V4
16
V19
24
V57
29
V15
≤ 27
> 27
≤ 36
> 36
≤ 36
> 36
≤ 49
> 49
≤ 31
> 31
5
V19
10
V51
17
V15
≤ 35
> 35
≤ 25
> 25
≤ 36
> 36
Node 6 (n = 18
Node 7 (n = 1
Node 8 (n = 21
Node 11 (n = 1
Node 12 (n = 1
Node 13 (n = 4
Node 15 (n = 3
Node 18 (n = 5
Node 19 (n = 1
Node 20 (n = 1
Node 23 (n = 1
Node 25 (n = 2
Node 26 (n = 1
Node 28 (n = 1
Node 30 (n = 1
Node 31 (n = 55)
1
0.8
0.6
0.4
0.2
0
14

REFERENCES:

- Aggarwal, C.C. (2015). *Data Mining: The Textbook*. New York: Springer.
- Bowerman, B.L., O'Connel, R.T., Koehler, A.B. (2005). *Forecasting, time series, and regression: an applied approach. 4th edition*. Belmont: thompson Learning.
- Chatfield, C., Xing, H. (2019). *The Analysis Of Time Series: An Introduction with R*. Taylor and Francis.
- Maharaj, E.N., D'Urso, P., Caiado, J. (2019). *Time Series Clustering and Classification*. Chapman and Hall
- Montero, P., Vilar, J.A. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software* 62 (1): 1-43.
- Sardá-Espinosa, A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal* (11/01): 1-22.
- Shumway, R., Stoffer, D. (2019). *Time Series: A Data Analysis Approach Using R*. CRC Press
- Woodward, W.A., Gray, H.L., Elliott, A.C. (2021). *Applied Time Series Analysis with R. 2nd edition*. CRC Press.

NEXT TOPIC:

Mining Sequence Data