Faculty of Cognitive Sciences and Human Development

IMPLEMENTATION OF A TIME SERIES PREDICTION ALGORITHM FOR COVID-19 CONFIRMED CASES IN MALAYSIA: A PRELIMANARY STUDY AFTER THE RECOVERY MOVEMENT CONTROL ORDER

Tee Bee Lin

Bachelor of Science with Honours
(Cognitive Science)
2020

UNIVERSITI MALAYSIA SARAWAK

Grade: A-

Please tick one

Final Year Project Report ☒

Masters ☐

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 7 day of AUGUST year 2020.

Student's Declaration:

I, TEE BEE LIN, 62754, FACULTY OF COGNITIVE SCIENCES AND HUMAN DEVELOPMENT, hereby declare that the work entitled, IMPLEMENTATION OF A TIME SERIES PREDICTION ALGORITHM FOR COVID-19 CONFIRMED CASES IN MALAYSIA: A PRELIMANARY STUDY AFTER THE RECOVERY MOVEMENT CONTROL ORDER is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

7 AUGUST 2020

_____

_____

TEE BEE LIN (62754)

Supervisor's Declaration:

I, ASSOC PROF DR TEH CHEE SIONG , hereby certify that the work entitled, IMPLEMENTATION OF A TIME SERIES PREDICTION ALGORITHM FOR COVID-19 CONFIRMED CASES IN MALAYSIA: A PRELIMANARY STUDY AFTER THE RECOVERY MOVEMENT CONTROL ORDER was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a *partial/full fulfillment for the conferment of BACHELOR OF SCIENCE WITH HONOURS (COGNITIVE SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work.

Received for examination by: TC _____ Date: 7 AUGUST 2020

(ASSOC PROF DR TEH CHEE SIONG)

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*****
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*****
☒ **OPEN ACCESS**

I declare this Project/~~Thesis~~ is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (√)):

☒ **YES**
☐ **NO**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature: _____ Supervisor's signature: _____
Date: 7 August 2020 Date: 7 August 2020

Current Address:

Notes: ***** If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

IMPLEMENTATION OF A TIME SERIES PREDICTION ALGORITHM FOR COVID-19 CONFIRMED CASES IN MALAYSIA: A PRELIMANARY STUDY AFTER THE RECOVERY MOVEMENT CONTROL ORDER

TEE BEE LIN

This project is submitted
in partial fulfilment of the requirement for a
Bachelor of Science with Honours
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
(2020)

The project entitled 'IMPLEMENTATION OF A TIME SERIES PREDICTION ALGORITHM FOR COVID-19 CONFIRMED CASES IN MALAYSIA: A PRELIMANARY STUDY AFTER THE RECOVERY MOVEMENT CONTROL ORDER' was prepared by Tee Bee Lin and submitted to the Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours (Cognitive Science).

Received for examination by:

TC

(ASSOC. PROF. DR TEH CHEE SIONG)

Date:
7 AUGUST 2020

Grade
A-

ACKNOWLEDGEMENTS

I would like to address my deep gratitude to my supervisor, Associate Professor Dr Teh Chee Siong for his patient guidance of this research and assistance in keeping my progress on schedule. I am very grateful to his supervision and enthusiastic encouragement on every step along the path of completing this project.

Next, I wish to extend my appreciation to my fellow peers: Ong Hui Xin, Goh Cheng Lynn, Maxqueen anak Kennedy, Lim Yi Ann, Ling Chien and Yee Chee Ling for their motivation and support while doing this project.

Finally, special thanks to my mother who supports me financially and emotionally throughout my entire study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Time series analysis is a method to predict future values by reading the previous data in time format. It is widely used nowadays especially weather forecast, economy trading, many industries are utilizing it for their business planning, etc. Meanwhile, coronavirus disease 2019, also known as COVID-19, is the pandemic that has been researched extensively around the world. In Malaysia, the Recovery MCO (RMCO) is starting from 10th June 2020 until 31st August 2020 where certain social, educational and businesses are allowed to operate after the spread of disease starts decreasing (Loo, 2020). But there is new cluster reported on 16th July after the cases were thought decreasing gradually (Kaur, 2020). Although it is not announced as second outbreak yet it is necessary to study the time series Covid-19 cases in Malaysia to forecast the possible number of cases in future for better preparation and planning. In this research, ARIMA time series model applied to predict the Covid-19 future daily cases by adapting their time series dataset from the Johns Hopkins University official website. The data extracts Malaysia's Covid-19 confirmed cases from 22nd January 2020 to 30th July 2020. The parameter of ARIMA (2,1,1) has the highest accuracy compared to other tested parameters with 10.4269 of RMSE value and 2.1548 of MAPE value. The result of the predicted number of Covid-19 daily cases is between 15 to 17 in fourteen days with a 95% confidence interval.

Keywords: Time series analysis, COVID-19, ARIMA

ABSTRAK

Analisis siri masa adalah cara untuk meramalkan nilai masa depan dengan membace data sebelumny dalam format masa. Ia digunakan secara meluas pada masa kini terutamanya ramalan cuaca, perdagangan ekonomi dan pelbagai industry menggunakanny bagi perancangan perniagaannya dan lain-lain. Sementara itu, penyakit coronavirus 2019 (COVID-19) adalah pendemik yang banyak dikaji di seluruh dunia. Setelah Perintah Kawalan Pergerakan Pemulihan (PKPP) bermula dari 10 Jun 2020 hingga 31 Ogos 2020 di mana hanya perniagaan yang tertentu dibenarkan beroperasi apabila penyebaran penyakit mulai berkurang (Loo, 2020). Namun, terdapat kluster baru yang dilaporkan pada 16 Julai setelah kes-kes Covid-19 dianggap menurun (Kaur, 2020). Walaupun ia belum diumumkan sebagai wabak kedua, kajian siri masa dalam kes Covid-19 Malaysia perlu dilakukan untuk meramalkan jumlah kes di masa depan untuk bersedia and membuat perancangan yang lebih baik untuk mengelakkan perlakuan wabak kedua. Dalam penyelidikan tersebut, ARIMA model siri masa yang berlaku untuk meramalkan kes harian Covid-19 masa depan dengan mengambil set data siri masa dari laman web rasmi Johns Hopkins Universiti. Data mengekstrak kes-kes Covid-19 dalam Malaysia yang disahkan dari 22 Januari 2020 hingga 30 Julai 2020. Parameter ARIMA (2,1,1) mempunyai ketepatan tertinggi berbanding parameter lain yang diuji dengan 10.4269 dalam RMSE dan 2.1548 dalam MAPE. Hasil ramalan jumlah kes harian Covid-19 adalah antara 15 hingga 17 dalam empat belas hari dengan selang keyakinan 95%.

Kata Kunci: Time series analysis, COVID-19, ARIMA

CHAPTER 1

INTRODUCTION

Artificial Intelligence (AI), the term coined by John McCarthy, is the field that has been studied since the 40s including developing the intelligent system, machines, neural network, etc. by applying different principles from computer science, cognitive psychology, philosophy, linguistics and other related subjects (Russell & Novig, 2010). The further research of AI has extended into different subfield as the reliance on technology from society increasing, especially Data Science. **Data Science** is the field that generalizes the extraction of information from heterogeneous and unstructured data (Dhar, 2013). Data Science evolved from the study of pattern recognition which is the sets of data gathered is being analysed to discover patterns**.** It has emerged as trending and significant discipline in recent years which consist of multidisciplinary; those are **Data Analytics** and **Data Mining** (EDUCBA, n.d.a, n.d.b). Both disciplinaries are inevitable in business analytics, business intelligence and predictive analytics which are best known for time series prediction in many industries.

Meanwhile, **time series analysis** as the subdivide of data science using pattern recognition and prediction on the data collected. The time-series data are gathered over some period of regular or irregular time to develop a model which represents the built-in structure of the series. The function of the time series model is applied to generate the outcome for future prediction which is called forecasting. Forecasting is the process of statistical prediction widely used in various industries. It can be demanded in various conditions, schedule numbers of workers in train station counter next week by predicting the number passengers; managing expenses for the company that sells certain products by the forecast of sales; determine on whether building another power generation plant in a few years by predicting upcoming request.

It also can be involved many years earlier for capital expenditure cases or couple of minutes in advance for telecommunication networks. It is an essential planning aid whatever the conditions or time horizons required. There are three factors to be considered for the predictability of a number or an incident. One, the understanding of factors that contributed. Two, the numbers of the available data. Three, whether the result to be forecast can be affected by the forecasts. For instance, electricity demand forecasting can be distinctly precise due to mostly fulfilling of the three conditions (Hyndman & Athanasopoulos, 2018).

1.1 Research Background

Time series analytics contributes greatly nowadays, especially in health care. Recently, COVID-19 known as the global public health crisis, many kinds of research and industries using the time series analysis to forecast the health pattern, employment rate, economy rate, number of sales in future months, etc. to understand and prepare for the possible outcomes, so that they can overcome worst circumstances. COVID-19, a latest discovered coronavirus strain which highly infectious respiratory disease that was originated in December 2019 from Wuhan, China (WHO, n.d.a). Based on the World Health Organization (WHO), the total number of Covid-19 confirm cases was 17,005,893 worldwide while Malaysia had a total of 8,964 cases on 30th July 2020 (WHO, n.d.b).

There are many time series models are used in the researches such as classical mathematical model: ARMA, ARIMA, VAR, Holt-Winter Exponential, Linear Regression, GARCH, etc.; neural network model: RNN, LSTM, GRU, ANFIS, Transformer, etc. Both classical mathematical model and the neural network model have different properties. Classical time series prediction models are most suitable for univariate time series data and they are come

in handy when the size of the data is not large, while neural network models are vice versa. Neural network models can predict a large amount of data and multiple variables at the same time.

In this research, the time series dataset that will be working on is the ongoing COVID-19 pandemic has a total of 190 days from 22nd January 2020 to 30th July 2020. This means that the dataset is not large and ARIMA will be used for this prediction. This is because ARIMA (Autoregressive Integrated Moving Average) is a popular and widely used time series prediction methods in analysing stationary univariate time series data. This is because of the simplicity and systematic structure of the model and its acceptable forecasting performance (Wang et al., 2018).

The ARIMA model will be applied in this experiment and discussed in Chapter 2. In this study, the algorithms will be used for analysing the valid datasets for accuracy by comparing different parameter. The procedure of this research will discuss in Chapter 3.

1.2 Problem Statements and Motivations

Time series forecasting is one of the methods in predictive modelling that solves prediction besides of classification, clustering and others. It plays an important role in data science as the time component is mostly the issue when making predictions. The time series forecasting model is determined on the skill by the performance at predicting future data. Thus, the reason for specific prediction was made for better explaining and a greater understanding of the underlying causes behind the problem (Brownlee, 2016).

Research on time series analysis are studied over decades and various mathematical models have developed for time series prediction. However, the knowledge of time series is

often neglected due to its complexity and it is important to be the focus on the upcoming new era. The application of time series can be seen in various contexts such as daily weather temperature, allocation of resources, business planning, cryptocurrency, public health pattern and stock price forecasting (Erica, 2019). We can see that it is significantly involved in our daily life and it would be benefitting a lot to people to understand it better.

On the other side, Covid-19 as a worldwide pandemic has been studied globally especially China, US and other seriously infected countries. Malaysia time series cases have been less studied on because the situation seems controlled and movement control order (MCO) is extended into Conditional MCO (CMCO) and Recovery MCO (RMCO) where most businesses are allowed to open. Yet, the precautions of personal hygiene and practise social distancing are required to avoid the next outbreak. The study of Covid-19 in Malaysia is needed to forecast the possible case in future if we follow the policies given by the government.

Different types of time series forecasting models are necessary to be examined and the algorithm and performance in time series prediction are needed to be studied. This research also reviews some of the existing approaches and algorithms for forecasting time series to better understand the performances.

1.3 Research Objectives

1.3.1 General Objective

Generally, this study aims to conduct experimental studies by implementing time series model for COVID-19 datasets.

1.3.2 Specific Objective

a) To analyse and study in depth of the ARIMA model and related time series prediction algorithms.
b) To experiment the selected model using simulated and well-known database which is the COVID-19 datasets from the official site of Johns Hopkins University.
c) To determine the parameters based on the prediction accuracy of the algorithms used.

1.4 Research Scope

This research is focused on compilation and experimentation of the mathematical model to test using a well-known database.

1.5 Significance of the Study

The finding of this study is to contribute the knowledge of big data practice and theory by comparing and evaluating the accuracy of the varied parameter of the prediction models. The latest deep learning algorithm will be further explored as future research.

CHAPTER 2

LITERATURE REVIEW

2.0 Overview

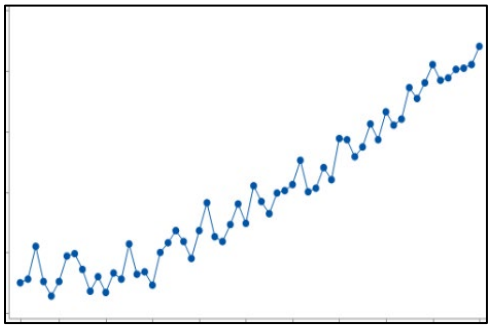
This chapter explains about the depth of what is time series prediction and types of data can be applied. The models of the time series such as ARIMA and the latest related research also will be discussed.

2.1 Time Series forecasting

Time series is a type of discrete stochastic process while stationary is the primary assumption to influence its estimation to structural parameters (Zhang et al., 2017). It is mean that a set of vectors $y(t)$, where the variable $y(t)$ is considered as random, where t stands for the time elapsed which equals to 0, 1, 2, …, n (Mohamed El-Hawary, 2017).

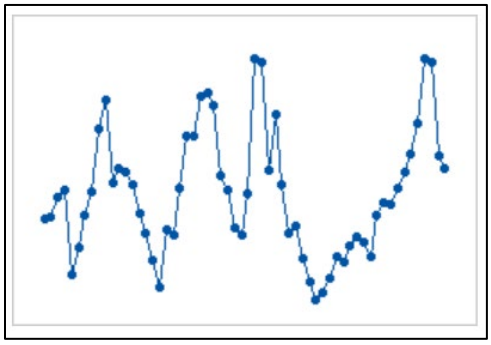
During an activity, the measurements can be taken in a time series and organized in chronological order. Univariate is a term that a time series comprises a single variable in the records while multivariate means the it has more than one variable in the records. The observations of discrete time series are measured with time intervals data points such as company production, the exchange rate between currencies and city population. However, the observations of continuous time series are measured at random time data that requires constantly recorded like temperature readings, earthquakes and velocity of an airplane (Senter, 2014). There are four main components that can affect the observed data which are *Trend*, *Seasonal*, *Cyclical* and *Irregular* (Adhikari & Agrawal, 2013).

Trend



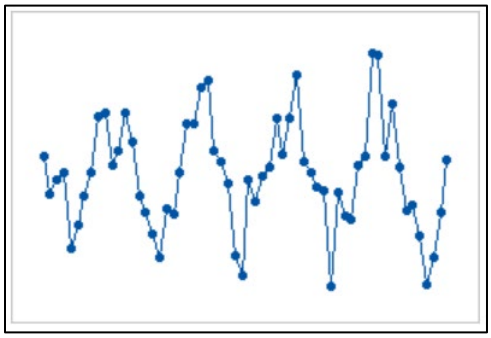
Secular Trend defined as the general tendency of a time series to increase (upward trend), decrease (downward trend) or stagnate over a long period of time. Trend is a smooth and long-term movement in a time series. For instance, it can be a series relating to number of cars or the population growth in a country, mortality rates, etc.

Cyclical



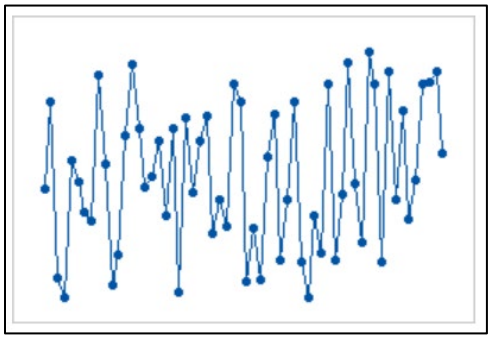
Cyclical variation represents the events repeat in cycles caused medium-term changes in the series. The cycle duration extends over a longer time period, generally more than two years. Most of the financial and economic time series show some cyclical variation.

Seasonal



Seasonal variations are usually fluctuating within a year during the season. The significant factors that cause seasonal variations are: weather and climate conditions, customs, traditional habits, etc. For instance, during winter, the woollen cloth sales increases, while during summer, ice cream sales increases. Seasonal variation is an important factor especially for entrepreneur, manufacturer and shopkeeper for upcoming business plans.

Irregular/Residuals



Random variations are not regular and the patterns are not repeating, these are affected by unpredictable influences events such as flood, earthquake, revolution, strike, war, etc.

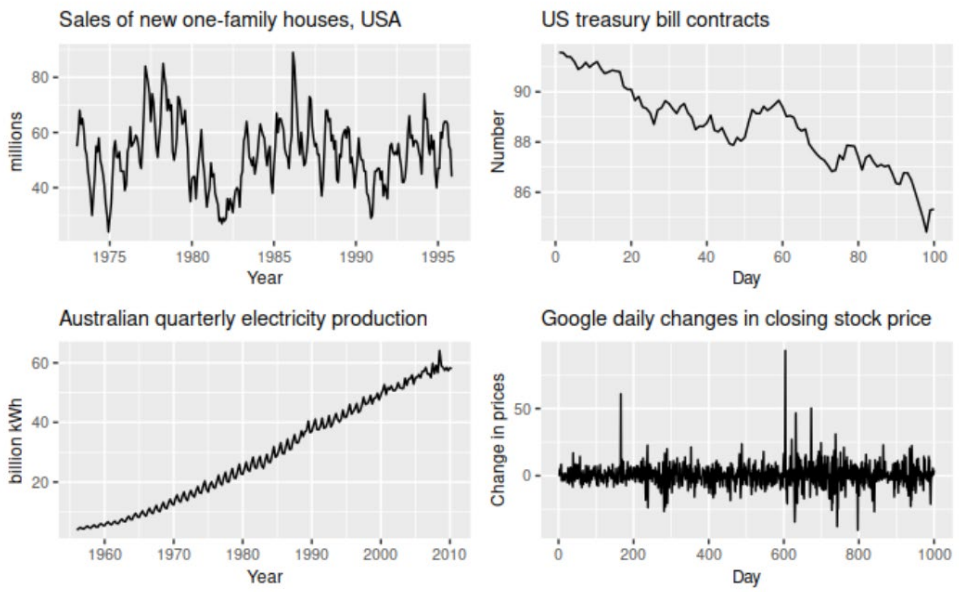


Figure 1. The examples of time series demonstrating different patterns. Adapted from *Forecasting: Principles and Practice*, by R. J. Hyndman & G. Athanasopoulos, 2018, MEL, AU: OTexts.

a. The graph at the top left represents seasonality within each year with some strong cyclic behavior in a period of between six to ten years and there is no obvious trend in the data over this period.
b. The graph at the top right shows 100 consecutive trading days of a market. There is a downward trend but no signs of seasonality. If there is a longer series, the downward trend may view as part of a long cycle, but it appears as trend because it only more than 100 days are viewed.
c. The graph at the bottom left represents an increasing trend with strong seasonality.
d. The graph at the bottom right shows no trend, cyclic or seasonality behavior. it is irregular or residual since There are random fluctuations that appears to be not easily to predict.

2.1.1 Stationary

Stationarity is the result of showing constant statistical properties such as mean, variance and co-variance;

Condition 1: Mean of the time series must be a constant

Condition 2: Variance of the series must be a constant

Condition 3: Covariance of the *i* th term and the (*i* + *m*)th term should not be a function of time, also means should not have the sign of seasonality

The points below show the method to check the stationarity of the graph/chart

1. **Visually**

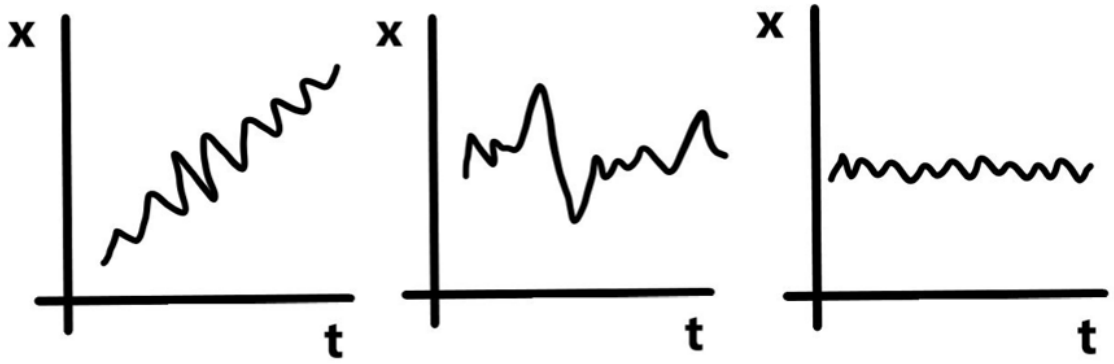


Figure 2. The non-stationary time series

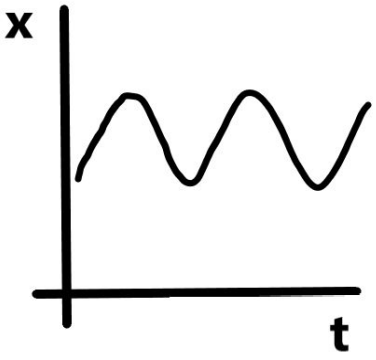


Figure 3. The stationary time series

Based on *Figure 2*, the first graph does not fulfil the first condition which mean is not constant while the variance of the second graph is not constant and the spread of time in the last graph is increasing so it does not fulfil the third condition.

2. Augmented Dicky-Fuller (ADF) test

According to Holmes et. al (2019), the ADF test is a standard statistical significance test used to analyse the stationarity of a time series. It has a hypothesis testing consist of a null (H_0) and alternate hypothesis (H_1), where p-value > 0.05 is accepted null hypothesis whereas p-value < 0.05 is vice versa. The equation of the ADF test is shown below (Prabhakaran, n.d.; Holmes et. al, 2019):

$$y_t = c + \beta_t + \alpha y_{t-1} + \emptyset_1 \Delta Y_{t-1} + \emptyset_2 \Delta Y_{t-2} + \cdots + \emptyset_p \Delta Y_{t-p} + e_t \quad \textbf{(1)}$$

where,

y_{t-1} : lag 1 of time series

ΔY_{t-1} : first differencing of the time series

When the α=1, it contains the presence of unit root, the p-value obtained is more than 0.05 significance level which considers that the null hypothesis is accepted. Thus, it refers that the series is non-stationary. To identify a stationary series, the p-value has to be less than 0.05 to accept the alternative hypothesis and reject null hypothesis.

2.1.2 Techniques to make the non-stationary graph/chart to stationary

Differencing

It is the method for transforming time series data. There is another series transforming technique called de-trending is applied to removing linear or nonlinear trend by deducting the

trend line with the plot and this kind of time series is known as trend-stationary. However, this formula is not adequate in many cases for achieving stationarity, especially for stochastic trend. Therefore, another way to transform a series is to apply differencing. The equation of differencing is shown below (Gupta et al., 2018; Hyndman & Anthanasopoulos, 2018):

First-order differencing is used which changing from a period to the next period.

$$z'_{t_{new}} = z_{t_{initial}} - z_{t-m} \quad \textbf{(2)}$$

where,

$z'_{t_{new}}$: new observation value

$z_{t_{initial}}$: previous observation value before differencing,

z_{t-m} : time period t,

m: number/duration of seasons or lag differences,

*Note that $m = 1$ is used as the first order differencing method.

For second-order differencing, substitute $m = 1$,

$$\begin{aligned} z''_{t_{new}} &= z'_{t_{initial}} - z'_{t-1} \quad \textbf{(3)} \\ &= (z_t - z_{t-1}) - (z_{t-1} - z_{t-2}) \\ &= z_t - 2z_{t-1} + z_{t-2} \end{aligned}$$

where,

$z''_{t_{new}}$: new observation value

$z'_{t_{initial}}$: previous observation value before differencing,

z'_{t-1} : time period t,