NUMBER OF COVID-19 INFECTION CASES FORECAST IN MALAYSIA

LEW TECK WEI

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

NUMBER OF COVID-19 INFECTION CASES FORECAST IN MALAYSIA

LEW TECK WEI

THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF DATA SCIENCE

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: LEW TECK WEI (I.C/Passport No: 891005-14-6335)

Matric No: 70216821

Name of Degree: Master of Data Science

Number of Covid-19 Infection Cases Forecast In Malaysia:

Field of Study: Public Health Care

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 01/01/2021

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

[NUMBER OF COVID-19 INFECTION CASES FORECAST IN MALAYSIA]

ABSTRACT

The recent Covid-19 pandemic has impacted our daily' living and there are number of measures are being implemented to control the spread of the contagious. Considering the economic impacts also, government and the public have been putting lax on the controls and the number of covid-19 infection cases has seen dramatically increased. Great worry has always been stayed on the healthcare facilities are sufficiently to cater the impending outbreak. This paper is to conduct an experiment to forecast the number of covid-19 infection cases using time series model. The model suggest that the number of daily covid-19 infection cases is on an uptrend with a wider region between maximum and minimum threshold.

Keywords: Covid-19, ARIMA, Detrending

TABLE OF CONTENTS

[NUMBER OF COVID-19 INFECTION CASES FORECAST IN MALAYSIA]

CHAPTER 1:

1.1 Introduction

The global outbreak of covid-19 pandemic has exhausted every country's economy severely and some countries has not even implemented the movement control operations (MCO) as Malaysia does where all the businesses were mandatorily shut down. The exponentially increased number of covid-19 infection cases can be witnessed especially in western countries where the government has put higher weight on economic consideration than controlling the spread of viruses. In contrast, there are several measures implemented by many Asian countries especially in Malaysia, the number of covid-19 infection cases had successfully been constrained to not-more-than 3 digits cases. Sadly, the measures have been taking toll on the country's economy and the public was pushed into living struggle despite the effectiveness. Society insecurity was also brought by public on the empathy side that whether those living hand to mouth or the poverty group would resort to crime to earn a living for their family or various crimes including violent ones during this pandemic implementation that bring pressure.[1] On the contrary, the health care facilities were seen unready to accommodate the sudden outbreak leading an unfortunate implementation of MCO. Ever since the great impact on social and economic that felt, the government had shifted from a stricter control namely MCO to lenient one namely Recovery Movement Control Operations (RMCO) / Conditional Movement Control Operations (CMCO) at the price of expected increase in the number of covid-19 infection cases. Putting lax on the controls equally to what the western

[1] Hassanal Noor Rashid (2020, April 7). "COVID-19 crisis: Crime, poverty and the need to evolve social security." Astro Aswana, Retrieved from: https://www.astroawani.com/berita-malaysia/covid19-crisis-crime-poverty-and-the-need-to-evolve-social-security-237089

government has situated their country where number of infection and death case are of catastrophic. The sacrifice has to be made for any option chosen at the minimal impact to the overall picture. However, evidenced from the extremely alarming daily number of infection/ death cases reported in western counties (e.g. US), a question raised among the public on the readiness and adequacy of healthcare facilities amid the expected daily high increased of infection cases.

1.2 Objective

The objective of the experiment / study is to predict the expected number of infection cases based on the historical daily number of infection cases reported for the past one years. Moreover, the prediction will be used as a reference/ indicator for the government to plan or to allocate the budget of health care facilities to ensure sufficient health care facilities. Availability of health care facilities was highlighted in a report published by A the Pharmaceutical Services Programme under Ministry of Health Malaysia in 2020 stating that the supply chain management of medicine and related consumables stockpile was being monitored especially for the high usage items to ensure no stock disruptions.[2] In view of this, the objective of this study is to understand 3 research questions as below:

a) What is the expected daily number of Covid-19 infection cases between the maximum and minimal region?

b) Does the current healthcare facility including the health care work force sufficient to accommodate the expected increased number of Covid-19 infection cases?

[2] Pharmaceutical Services Programme. (2020). "COVID-19 Pandemic in Malaysia: The Journey". A Report by the Pharmaceutical Services Programme, Ministry of Health Malaysia, p.6

c) How much budget should be ready to cater the expected increased number of Covid- 19 infection cases?

1.3 Description of Dataset

The source of the dataset is from EU Open Data Portal (or European Union Open Data). The dataset has 12 columns and 59,983 rows which generally contains the statistic of covid-19 infection and death cases worldwide as described below:

Features/ Column	**Description of Features**
dateRep	Date
day	Day of the Date
month	Month of the Date
year	Year of the Date
cases	Cases of the Day
deaths	Deaths of the Day
countriesAndTerritories	Countries
geoId	Geography ID
countryterritoryCode	Country territory Code
popData2019	Population
continentExp	Continent
Cumulative_number_for_14_days_of_COVID-19_cases_per_100000	Cumulative Number for 14-days of Covid-19 infection cases

Table 1: Description of Dataset

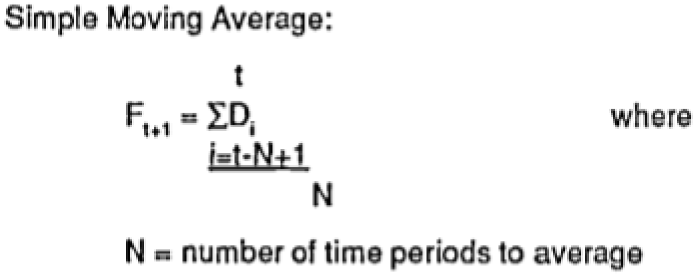
CHAPTER 2:

2.1 Analysis

This analysis involves the prediction using the time series dataset prescribed in chapter 1.3. The machine learning tool that is employed in this experiment is time-series forecast. Time series forecast is used to predict the coming series of periodic observations of a quantitative data based only on historical observations to forecast occurrences in the coming days[3].

2.1.1 De-trending

The times series model developed using simple moving average (SMA) to smooth out short-term deviations of time series data and indicate long-term trends or cycles[4]. With this method, the number of covid-19 infection cases of the next observation period by averaging in order to form an equally weighted of each past number of case of the day. The formulas as below:



Simple Moving Average:

$$F_{t+1} = \frac{\sum_{i=t-N+1}^{t} D_i}{N} \qquad \text{where}$$

N = number of time periods to average

1

A critical part of time-series forecasting model is de-trending. Detrending the time series is a critical part in forming a basis to estimate the business cycles/ cyclical pattern of variable.

[3] Samuel E. Bodily. (2008). "Time-Series Forecasting.". A*rticle of University of Virginia Darden School Foundation, Charlottesville, VA*. p.1

[4] C. L. Karmaker, P. K. Halder and E. Sarker. (2017). "A Study of Time Series Model for Predicting Jute Yarn Demand: Case Study". *Journal of Industrial Engineering, p. 2*

Detrending shows a different aspect of time series data by removing deterministic and stochastic trends from the SMA. [5]

Auto-regressive (AR) learns the behavioral pattern of the historical data to perform the time series forecasting to the prediction.

2.1.2 Detecting the MA and AR using ACF and PACF

2.1.2.1 Auto-correction Function (ACF or q)

ACF is employed to observe the correlation between the observation at current spot and observation at previous spot. In other meaning, it represents the correlation N between the residuals at their associated time t and those same residuals shifted ahead by one unit of time and the formulas as described below:

$$r_1 = \frac{\sum_{t=2}^{N} (e_t)(e_{t-1})}{\sum_{t=1}^{N} e_t^2},$$ [6]

coefficient is where et is the residual (i.e., the estimate of the error of the model), measured at time t, and N is the number of residuals in the observed time series.

[5] Viorica Chirila. (2015). “Detrending Time Series and Business Cycles.The Romanian Case”. Acta Universitastis Danubius, Vol, 8. No. 4. p. 135

[6] Bradley E. Huitema and Sean Laraway. (2006). "Autocorrelation." Encyclopedia of Measurement and Statistics. p.2

2.1.2.2 Partial Auto-correction Function (PACF or p)

PACF is used to remove the influence of the day before yesterday to observe the correlation between two time spots given that we consider both observation is correlated to other time spots. In other words, partial autocorrelation gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. Formulas as follow:

$$\boldsymbol{PACF(T_i\,, k = 2)} = \frac{\boldsymbol{Covariance}\ \ (T_i \mid T_{(i-1)}\ , T_{(i-2)} \mid T_{(i-1)})}{\sigma_{T_i \mid T_{(i-1)}} \times \sigma_{T_{(i-2)} \mid T_{(i-1)}}}$$

2.1.2.3 ARIMA

An ARIMAX model is as a multiple regression model with one or more autoregressive (AR) terms and/or one or more moving average (MA) terms and the formulas as below:

$$\hat{y}_t = \hat{\beta}_1 x_t + \hat{\phi}_1 y_{t-1} + \hat{\phi}_2 y_{t-2} + \hat{\theta}_1 \epsilon_{t-4},$$

where $\hat{\beta}_1$, $\hat{\phi}_1$, $\hat{\phi}_2$ and $\hat{\theta}_1$ are estimated coefficients. [7]

x_t= the value of the independent variable at time
Y_{t-1} = the immediately preceding value of the dependent variable at time

[7] Andrews, Bruce H. & Dean , Matthew D. & Swain , Robert & Cole ,Caroline. (2013). *"Building ARIMA and ARIMAX Models for Predicting Long-Term Disability Benefit Application Rates in the Public/Private Sector"*. University of Southern Maine. Society of Actuaries, p.7

Y_{t-2} = the immediately preceding value of the dependent variable at time
ϵ_{t-1}= the estimation error produced by the model at time

An autoregressive integrated moving average (ARIMA) model is employed to generalize the ARMA model. Thus, ARMA are fitted to the time serious data as to predict future trend.

The *"I"* represents a differencing step to eliminate the non-stationary of the mean function in the combination between AR and MA. It also acts as an addition or one more time of de-trending work mentioned in section 2.1.1. The ARIMA model is famous as the Box- Jenkins. ARIMA model can be employed for time series fitting and forecasting with a temporal correlation as it incorporates (d) differencing into the ARMA model and which was also designed for stationary time series[8] (Wong W.M (2020)

An ARIMA model is labeled as: ARIMA model (p, d, q), [9]

wherein:
p is the number of AR terms
d is the number of differences; and
q is the number of MA.

[8] Wong W.M. (2020). "Flood Prediction using ARIMA Model in Sungai Melaka Malaysia." *International Journal of Advanced Trends in Computer Science and Engineering*. Vol. 9, No. 4, p.5287

[9] Fattah, Jamal, et al. (2018). "Forecasting of demand using ARIMA model." *International Journal of Engineering Business Management 10*. p. 3

2.2 Design

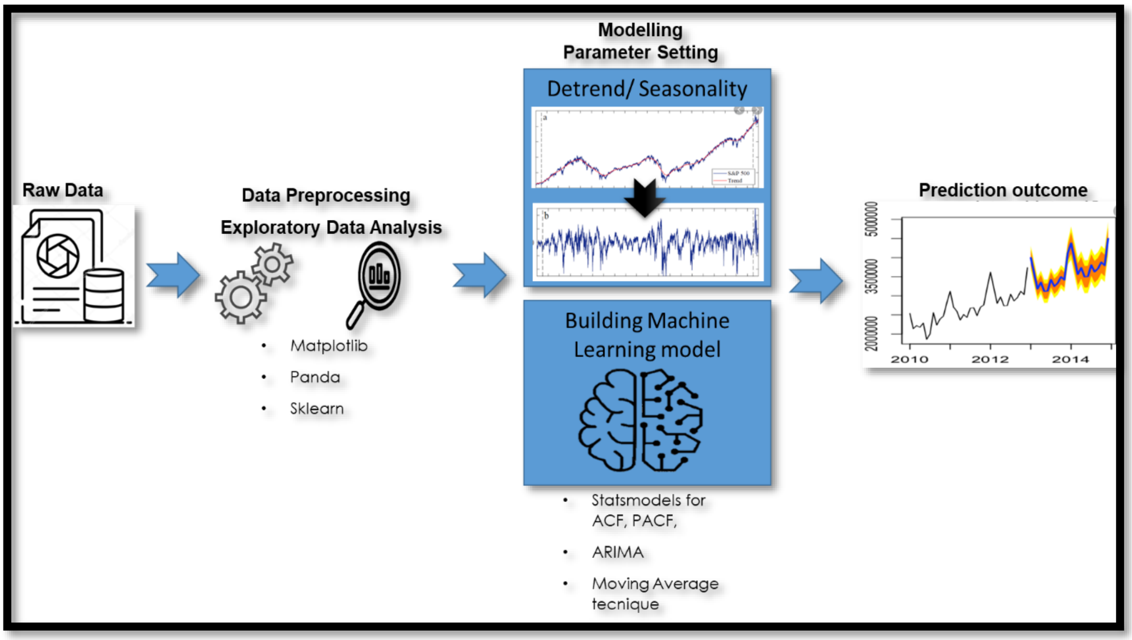
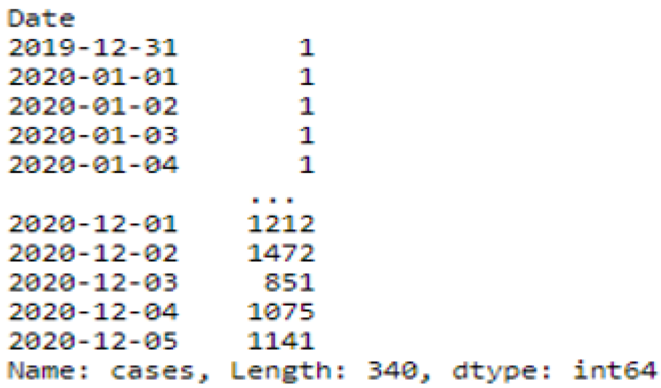


Chart 1: Research Design

2.2.1 Data preprocessing and Exploratory Data Analysis (EDA)

First stage involves removing unnecessary feature. Data removal (removing NAN Data) and filtering were conducted to remains the required features in this experiment. They are dateRep, cases of countriesAndTerrotories = MY. Apart from that, as it's a time series forecast model, data formatting was performed on the dateRep to convert the date into datetime format and was made to be an "index" data. The outcome of this stage as presented below:



```
Date
2019-12-31       1
2020-01-01       1
2020-01-02       1
2020-01-03       1
2020-01-04       1
              ...
2020-12-01    1212
2020-12-02    1472
2020-12-03     851
2020-12-04    1075
2020-12-05    1141
Name: cases, Length: 340, dtype: int64
```

Table 2: Dataset of number of reported daily covid-19 infection cases in Malayia

Some EDA works were carried out to familiar with the data as follows:

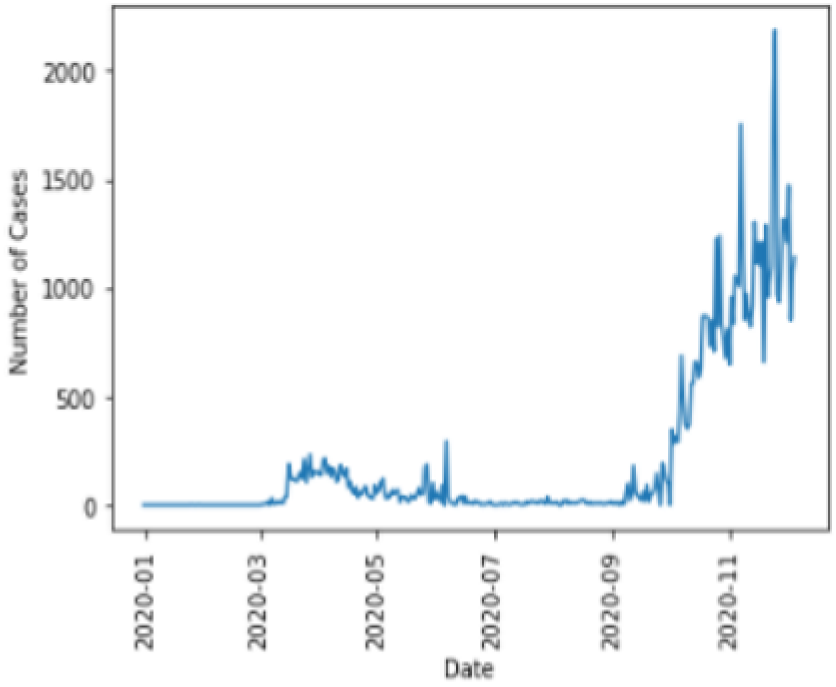


Chart 2: Line chart of the number of reported daily covid-19 infection cases in Malaysia

2.2.2 Building a Time-series Forecast Model

2.2.2.1 Data Normalization and Calculate 10-day SVM

Performance data normalization on feature “cases” within the scale of 0 to 1 and calculating the 10-days SVM. The result is shown by the below visualization:

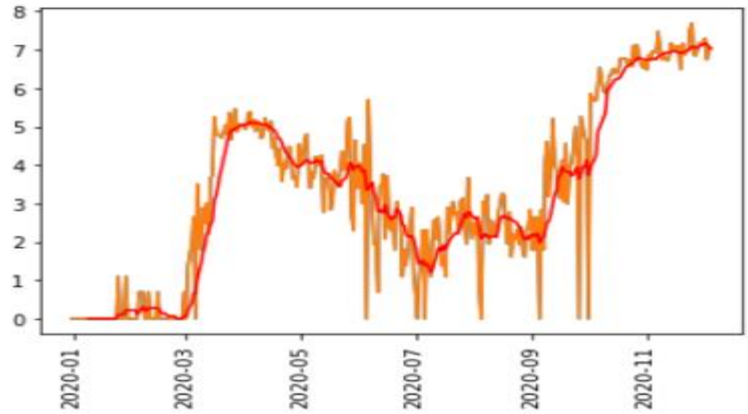


Chart 3: Line chart of the number of reported daily covid-19 infection cases in Malaysia (Normalized Data and 10-day SVM)

2.2.2.2 Removing the stationary

This was done by subtracting the shifted normalized data from the normalized data. The outcome of this work as presented below:

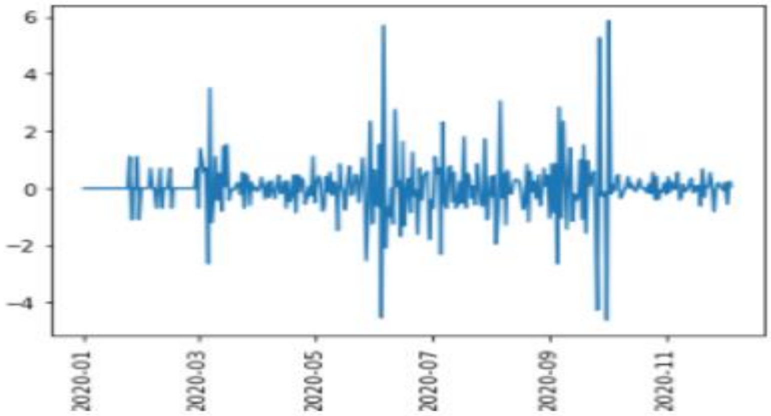


Chart 4: Line chart of the de-trended data for number of reported daily covid-19 infection cases in Malaysia

2.2.2.3 Removing the stationary

Checking to ensure de-trended data is of stationarity (as part of the rule for time-series forecast model). As can be seen from the below chart, the uptrend pattern7 as presented in “Chart 3: Line chart of the number of reported daily covid-19 infection cases in Malaysia (Normalized Data and 10-day SVM)" has been removed.

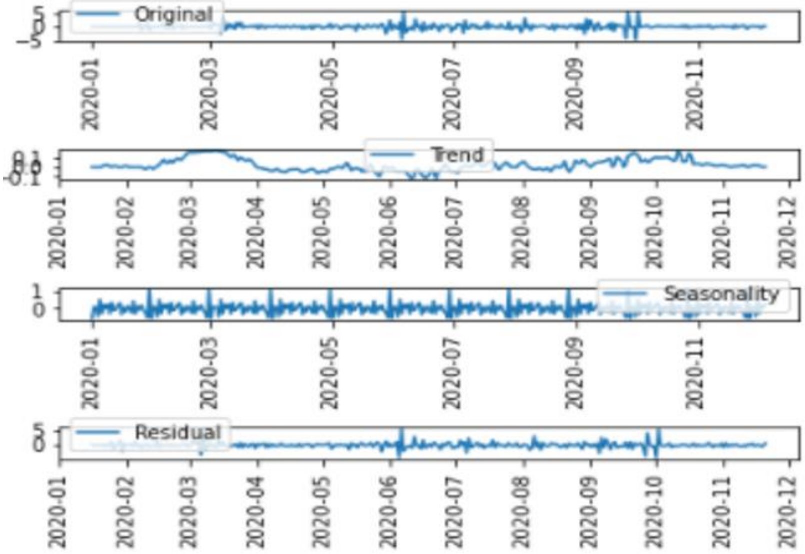


Chart 5: Line chart of the de-trended data for number of reported daily covid-19 infection cases in Malaysia

2.2.2.4 ACF (Auto-correction Function) & PACF (Partial Auto-Correlation Function)

To calculate the p and q value as part of ARIMA requirement, we have performed the calculation via the acf and pacf tools from statmodel package. As can be seen from the chart below, there is no correlation between the current spot and previous spot as presented by Autocorrelation function. Autocorrelation function is A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. Thus, we can conclude that the Moving Average (q) is equal to 0.

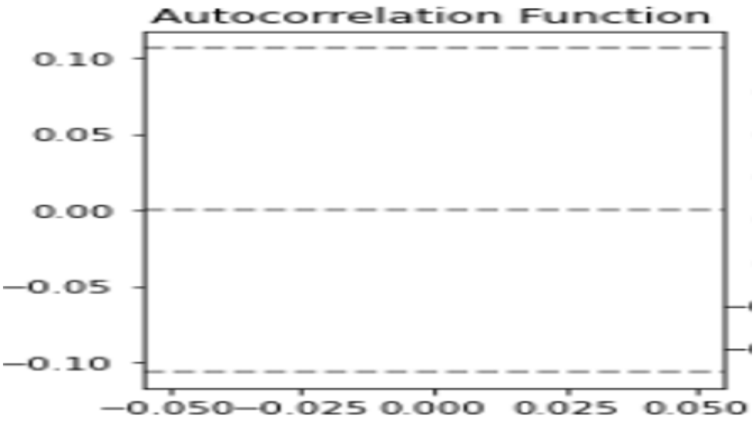


Chart 6: Autocorrelation Function Plot

As for the PACF plot below, it has a significant spike at lag1, meaning that higher-order partial autocorrelations are effectively explained by the lag1. Thus, the value of p equal to 1.

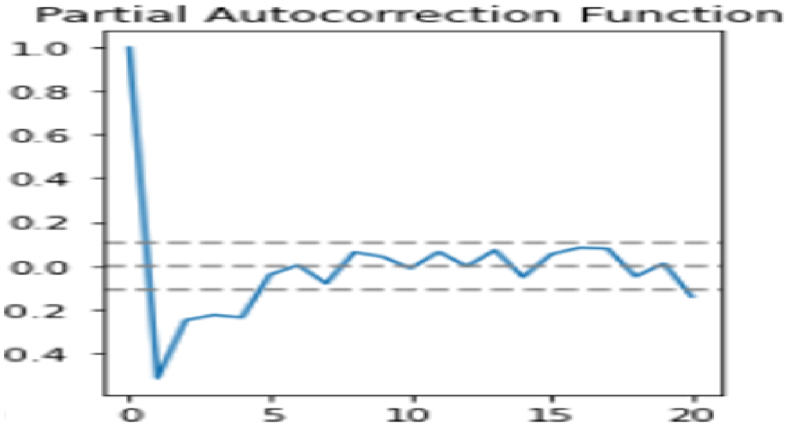


Chart 7: Autocorrelation Function Plot

2.2.2.5 ARIMA

Combining Auto Regression (AR) and Moving Average (MA) to build ARIMA to forecast the infection cases for the next year. The differing that is used in this experiment for “I” is 1 while q = 0 and p = 1 which is written as ARIMA (1,1,0).

CHAPTER 3: EXPERIMENT RESULTS

Based on the past one-year historical data. the ARIMA model has produced the prediction as shown in chart below. The Time-series forecasting shows that the number of cases is expected to be on uptrend with the highest and lowest number demonstrated in the grey region. This can be quite alarming as the number of cases can be as highest as 3500 cases per day. However, it can be observed that the number of minimum cases will also be on the uptrend with nearly close to a 1000 number of infection cases expected to be reported. Both maximum and minimum case are at the 95% confidence interval.

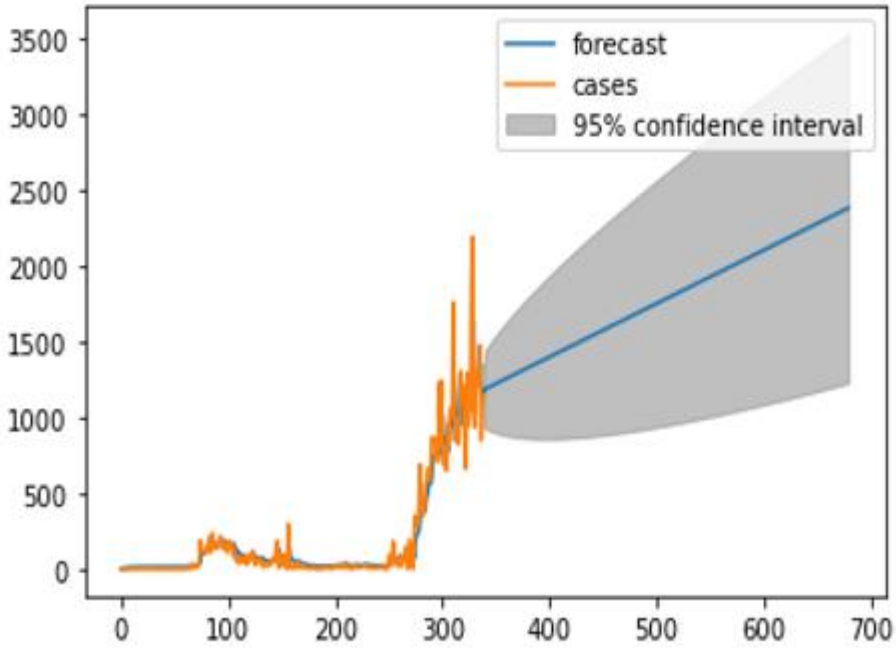


Chart 8: ARIMA plot

CHAPTER 4: DISCUSSION

ARIMA outputs forest case based on the previous value in the time-series (AR-term) and the error made by previous prediction. This typically allows the model to swiftly adjust for sudden changes to produce better prediction. The component of Auto regression (AR), Moving average (MA) and integrated (i) eliminate the common factor to make the models in a more parsimonious manner. [10] ARIMA models is backward looking, the long term forecast going to be straight line with poor prediction. The prediction ignores other factor such as the current movement control measure taken by the government to constraint the virus. In addition, the data in predicting the future cases is still not much to make this model perform as expected.

[10] Meyler, Aidan and Kenny, Geoff and Quinn, Terry. (1998). *"Forecasting irish inflation using ARIMA models"*. Central Bank and Financial Services Authority of Ireland, p.21

CHAPTER 5: CONCLUSION

The forecast by ARIMA models is still weak as the data is not much. In addition, the daily case between the days are upheaval, thus, the model many not be able to forecast with the right pattern and the seasonality. However, the model effectively given an uptrend with the expected daily case as high as 3500 cases. It is supported by the recent case where the number of reported case ranging from 2000 to 3500, which is in the region of the prediction shown in Chart 8: ARIMA plot.

LIST OF PUBLICATIONS/NEWS/ PAPERS PRESENTED

Primary Sources

Pharmaceutical Services Programme. (2020). “COVID-19 Pandemic in Malaysia: The Journey”. A Report by the Pharmaceutical Services Programme, Ministry of Health Malaysia.

Journal Article

Samuel E. Bodily. (2008). “Time-Series Forecasting.”. Article of University of Virginia Darden School Foundation, Charlottesville, VA.

Viorica Chirila. (2015). “Detrending Time Series and Business Cycles.The Romanian Case”. Acta Universitastis Danubius, Vol, 8. No. 4.

Bradley E. Huitema and Sean Laraway. (2006). "Autocorrelation." Encyclopedia of Measurement and Statistics.

Fattah, Jamal, et al. (2018). "Forecasting of demand using ARIMA model." *International Journal of Engineering Business Management 10*.

Wong W.M (2020). Flood Prediction using ARIMA Model in Sungai Melaka Malaysia. “International Journal of Advanced Trends in Computer Science and Engineering”. 9(4):5287 – 5295 DOI: 10.30534/ijatcse/2020/160942020. Retrieved from https://www.researchgate.net/publication/344297098_Flood_Prediction_using_ARIMA_Model_in_Sungai_Melaka_Malaysia/comments

Newspaper

Hassanal Noor Rashid (2020, April 7). COVID-19 crisis: Crime, poverty and the need to evolve social security. Astro Aswana, Retrieved from: https://www.astroawani.com/berita-malaysia/covid19-crisis-crime-poverty-and-the-need-to-evolve-social-security-237089 on 12 Jan 2021.