Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 18, 2023, page 181 – 194

Time Series Forecasting of Daily COVID-19 Cases in Malaysia

Yap Say Nee, Adina Najwa Kamarudin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: adina.najwa@utm.my

Abstract
A study was conducted in Malaysia to analyze COVID-19 trends for the period from March to September 2022, aiming to develop more accurate forecasting methods and guide the government's resource allocation. The study aimed to determine the most accurate forecasting model for daily COVID-19 cases for the next 14 days from 19 August to 1 September, 2022. Trend analysis showed a clear seasonal pattern in daily COVID-19 cases, indicating external factors' impact, such as the reopening of universities. Forecasting models, including SARIMA, exponential smoothing, and LSTM, were used to predict daily COVID-19 cases, and the Holt Winter exponential smoothing method was found to have the lowest RMSE and MAE values, making it the best forecasting model.

Keywords: COVID-19; Forecasting; SARIMA; Exponential Smoothing; LSTM

1. Introduction

COVID-19 is a novel coronavirus that was first identified in 2019 and has led to a global health emergency. The virus has caused over 1.2 million deaths globally, making it one of the deadliest years in recent history (World Health Organization, 2021a). Forecasting the spread of COVID-19 is crucial to understanding the potential impact of the virus and developing strategies to mitigate its spread. In Malaysia, the government has utilized time series forecasting to inform travel restrictions, contact tracing efforts, and other public health measures. The accuracy of forecasting is critical to ensure the most effective measures are implemented. However, there is a lack of research on predicting COVID-19 cases during periods of reduced movement restrictions. To address this gap, a study is being conducted in Malaysia to analyze COVID-19 trends from March to September 2022. The study aims to provide a better understanding of the virus's spread, develop more accurate forecasting methods, and guide the government's resource allocation. By predicting the extent of the COVID-19 pandemic, the consequences of the virus can be grasped and appropriate strategies can be implemented to reduce its transmission.

The focus of this research study was primarily centered around identifying the most effective forecasting model based on their performance or measurement accuracy. This was achieved by comparing several appropriate forecasting methods for predicting daily COVID-19 cases for two weeks from 19 August to 1 September 2022. Another key objective for this study was to analyze the trends evident in the daily COVID-19 case dataset, using the Minitab software. The trend analysis was segmented into two parts, with the first period covering the months of March and April 2022, while the second period covered 1 May to 18 August 2022. Furthermore, the study aimed to predict the daily COVID-19 cases for the next 14 days, using forecasting methods like SARIMA, exponential smoothing, and LSTM. Ultimately, the final objective was to identify the best performance-based forecasting model by comparing different forecasting methods specifically for daily cases of COVID-19 over a defined time period on 19 August and 1 September 2022.

The pandemic has caused widespread panic and a decrease in global trade, impacting economies around the world (Lone & Ahmad, 2020). Preventing the spread of COVID-19 is crucial, and staying informed about the virus and disease is the most effective way to do so. Vaccination is also an important step in prevention but forecasting future cases of COVID-19 is necessary as the virus has the potential to mutate and become resistant to the vaccine.

2. Materials and methods

Trend Analysis

The trend analysis for the data is partitioned into two sections. A trend analysis of daily COVID-19 cases was done for the first part from 1 March to 30 April 2022. For the second part of the analysis, the data was collected from 1 May to 18 August 2022, during which time the Malaysia's government had imposed a complete relaxation of the nation. This allowed for a more accurate analysis of the data, as there were no external factors influencing the results. Afterwards, the selected data was imported to Minitab Software for evaluating the trend analysis using time series plot.

Box-Jenkins (BJ) Model

The ARIMA or Box-Jenkins (BJ) model is a widely known and used method for statistical forecasting. Its main focus is on predicting future values of time series based on patterns established in past values. It derives its name from its components: Autoregressive (AR), Integrated (I), and Moving Average (MA). AR refers to the model's dependence on its own past values, I is used to make the time series stationary, and MA incorporates prediction based on the past observations' errors. This model is a more generalized form of the ARMA model, which is a linear regression model with ARMA errors.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) Model is a powerful tool in statistical forecasting that is specifically designed to analyze and interpret time series data that exhibit seasonal patterns. Time series data are characterized by observations that are made over a period of time that are not independent and are influenced by a range of different factors that can cause fluctuations in the data. The SARIMA model is a combination of two other statistical models; the Autoregressive Integrated Moving Average (ARIMA) model and the seasonal component of the Autoregressive Moving Average (ARMA) model. An abbreviated representation of the model is

$$SARIMA(p,d,q)(P,D,Q)_S$$

p = number of autoregressive terms;
d = number of differences; and
q = number of moving averages.
The SARIMA model is expressed mathematically as follows:

$$\emptyset_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\emptyset_Q(B^S)w_t$$

where,
$\emptyset_p(B)$ = Ordinary autoregressive component;
$\theta_q(B)$ = Ordinary moving average component;
$\Phi_P(B^S)$ = Seasonal autoregressive component;
$\emptyset_Q(B^S)$ = Seasonal moving average component;
$(1-B)^d$ = Ordinary difference component of order *d;*
$(1-B^S)^D$ = Seasonal different component of order *D;*
y_t = non-stationary time-series;
w_t = Gaussian white noise process.

Box-Jenkins (BJ) modeling involves following steps, firstly ensuring stationarity of the time series, then identifying necessary ARIMA components if seasonal data identifying SARIMA component. After that, estimating parameters using optimization techniques, follow by validating the model, and lastly forecasting future values based on historical data.

Exponential Smoothing

Exponential smoothing uses weights based on historical data and has three methods: single, double, and triple exponential smoothing. Single exponential smoothing uses one parameter, while double adds a second parameter to capture trends, and triple adds a third parameter to capture seasonality in the data.

Simple Exponential Smoothing

The Simple Exponential Smoothing Method is used to predict the future value of a time series that is not trending (stationary) and the mean of the time series original data, Y_t changes slowly over time.
The simple exponential smoothing model is as follows:

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

where,
F_{t+1}= forecasting value for the next period of *t;*
F_t= forecasting value for the period of *t;*
Y_t= observed value for the period of *t;*
α= smoothing constant with the interval of 0 to 1

Double Exponential Smoothing

Holt's Exponential Smoothing method is a more sophisticated version of the simple exponential smoothing. Adding the growth factor to the smoothing equation helped to more accurately predict the trend. In addition, it is also appropriate to use when a time series is increasing or decreasing at a fixed rate.
The Holt's exponential smoothing model is as follows:

$$F_{t+m} = L_t + mb_t$$

where,

level series: $L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1})$

trend estimate: $b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$

where,
α= smoothing constant with the interval of 0 to 1;
β= smoothing constant with the interval of 0 to 1;
m= future period to be forecast.
The procedures for the Holt's Exponential Smoothing Method are as follow:

For starter, obtain the initial estimates for L_0 and b_0 by setting $L_0 = Y_1$and b_0 that can choose one of the following formula:

$$b_0 = Y_2 - Y_1$$

$$b_0 = \frac{Y_4 - Y_1}{3}$$

$$b_0 = 0$$

Next, calculate F_t when *t*=1. After that, use the following formula to update the estimates L_t and b_t, with setting of α and β by any value from 0 to 1. Afterwards, find the best possible combination of α and β that will give the minimum value for SSE. Then, the forecasting is made m steps ahead at time *t*.

Holt-Winters Exponential Smoothing

The Holt-Winters Exponential Smoothing model is an improved version of the simple exponential smoothing model. It is most commonly used for data sets that show both trend and seasonality. It has three parameter models which are an extension of Holt's exponential smoothing. The additional equation is to adjust the model for the seasonal component.

Aside from that, there are two types of the Holt-Winters Exponential Smoothing which are Multiplicative Holt-Winters Method and Additive Holt-Winters Method.
The Holt-Winters Multiplicative Exponential Smoothing model is as follows:

$$F_{t+m} = (L_t + mb_t)S_{t+m-s}$$

where,
Level series: $L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$

Trend estimate: $b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$

Seasonality factor: $S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$
where,
α= smoothing constant with the interval of 0 to 1;
β= smoothing constant with the interval of 0 to 1;
γ= smoothing constant with the interval of 0 to 1;
S= seasonal number
The procedures for the Multiplicative Exponential Smoothing Method are as follow:

Firstly, obtain the initialize value for L_0, b_0 and $S_1, S_2, \dots, S_s$ by setting as follow:

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s)$$

$$b_s = \frac{1}{s}\left[\frac{(Y_{s+1} - Y_1) + (Y_{s+2} - Y_2) + \cdots + (Y_{s+s} - Y_s)}{s}\right]$$

$$S_1 = \frac{Y_1}{L_s}, S_2 = \frac{Y_2}{L_s}, \dots, S_s = \frac{Y_s}{L_s}$$

Then, $F_{t+1}$is calculated using the following equation:

$$F_{t+1} = (L_t + b_t)S_{t+1-s}$$

Afterwards, the estimated parameters L_t, b_t and S_t are updated by the following equation:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$$

Finally, a suitable combination of $\alpha, \beta,$ and γ is found to minimize the SSE value.
The Holt-Winters Additive Exponential Smoothing model is as follows:

$$F_{t+m} = L_t + mb_t + S_{t+m-s}$$

where,

$$L_t = \alpha(Y_t - S_{t-s} + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s}$$

where,

α= smoothing constant with the interval of 0 to 1;
β= smoothing constant with the interval of 0 to 1;
γ= smoothing constant with the interval of 0 to 1;
S= seasonal number

The procedures for the Additive Exponential Smoothing Method are as follow:
Firstly, obtain the initialize value for L_0, b_0 and $S_1, S_2, \dots, S_s$ by setting as follow:

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s)$$

$$b_s = \frac{1}{s}\left[\frac{(Y_{s+1} - Y_1) + (Y_{s+2} - Y_2) + \cdots + (Y_{s+s} - Y_s)}{s}\right]$$

$$S_1 = Y_1 - L_s, S_2 = Y_2 - L_s, \dots, S_s = Y_s - L_s$$

Then, F_{t+1} is calculated using the following equation:

$$F_{t+1} = L_t + b_t m + S_{t+m-s}$$

Afterwards, the estimated parameters L_t, b_t and S_t are updated by the following equation:

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1)}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma(Y_t - L_t + (1-\gamma)S_{t-s}$$

Finally, a suitable combination of $\alpha, \beta,$ and γ is found to minimize the SSE value.

Long-Short Term Memory (LSTM) Model

Long-short term memory is a type of machine learning algorithm that uses memory blocks to store information and improve predictions. These memory blocks have at least one memory cell and input and output ports that are controlled by signals. The forget gate was later added to allow for certain memories to be forgotten and ensure the cell can learn from new inputs.

In order to better understand the LSTM model, it is necessary to examine the steps of the model. The model consists of four main steps: inputting data, encoding the data, processing the data, and outputting the results. By understanding each of these steps, it is possible to get a better understanding of how the LSTM model works.
The LSTM model must first reset the output from the previous model at time *t* in order to create the network:

$$f_t = \sigma[W_f(h_{t-1}, x_t)]$$

where, f_t is forget function, σ is activation function, w is corresponding weight matrices, h_t is model output and x_t is model input.

After the model is created, the decision of what information should be stored should be made. This process is divided into two parts. The input gate layer is responsible for deciding which values to update. This is done by the sigmoid layer creating a vector that contains possible new values. After these processes are completed, the next step is to combine the two and update the input.

$$i_t = \sigma[W_f(h_{t-1}, x_t)]$$

$$\tilde{C}_t = tanh[W_c(h_{t-1}, x_t)]$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t$$

where,
The hyperbolic tangent function, $tanh$ is used to scale the values into the range of -1 to 1.

i_t = input function;
C_t = candidate vector.
The decision of the network's output is finally made. The result of the sigmoid gate is determined by the output.

$$o_t = \sigma[W_o(h_{t-1}, x_t)]$$

$$h_t = o_t tanh\,(C_t)$$

where,
o_t = sigmoid function output

The output of the model is a filtered version of what was inputted into the cell state. This is done in order to improve the model's performance. The cell state is constantly updated based on the inputs received. This ensures that the model only outputs the most relevant information.

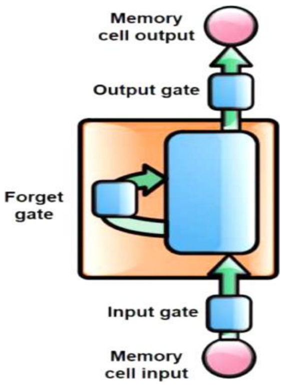

Figure 1 Internal architecture of LSTM (Source: Kırbaş et al. (2020))

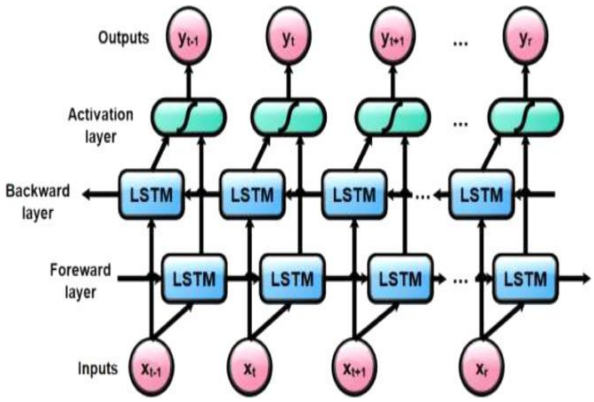

Figure 2 LSTM architecture (Source: Kırbaş et al. (2020))

The Long Short-Term Memory (LSTM) architecture is a type of recurrent neural network (RNN) that is designed to remember long-term dependencies. It is composed of four main components: the input gate, the forget gate, the output gate, and the memory cell. The input gate determines which part of the input is relevant and should be passed on to the memory cell. The forget gate determines which part of the memory cell should be forgotten or erased. The output gate determines which part of the memory cell should be passed on to the output. Finally, the memory cell stores the information for a long time and updates it based on the input and the forget gate. This architecture allows the network to learn long-term dependencies, making it suitable for complex problems.

Model Comparison

By comparing the values of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), the most suitable model can be identified.

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(y_t - F_t)^2}{n}}$$

$$MAE = \sum_{t=1}^{n} \frac{|y_t - F_t|}{n}$$

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|y_t - F_t|}{y_t}$$

where,
y_t = the original value of the data at time *t;*
F_t = the forecast value of the data at time *t;*
n = the number of data

These error metrics are commonly used to evaluate the accuracy of predictive models and can provide an indication of how well the model is able to make predictions. By comparing the errors of different models, a model with the lowest error can be identified as the most suitable one.

3. Results and discussion

Data Description

The purpose of this study is to predict the future prevalence of COVID-19. Data for this study were obtained from the WHO website which the data reported on Malaysia was provided by the Malaysian Ministry of Health. The website includes statistical reports such as daily cases and deaths by date reported to WHO and the latest reported counts of cases and deaths. However, the study only focused on daily cases of Covid-19 in Malaysia from 1 March 2022 to 1 September 2022. In the field of data analysis, the data are separated data into two distinct samples which are in-sample and out-sample data. 171 of the in-sample data is data that is used in the analysis itself and is the primary source of information for the study. Moreover, 14 of the out-sample data is data that is not used in the analysis, but instead serves as a comparison or benchmark to evaluate the accuracy of the analysis. By dividing data into these two distinct samples, it is better understanding on the data and draw more accurate conclusions.

Trend Analysis

Trend analysis is a process that involves analyzing the changes in data over a period of time. It helps to identify patterns or trends in data and can be used to make predictions about future behavior.

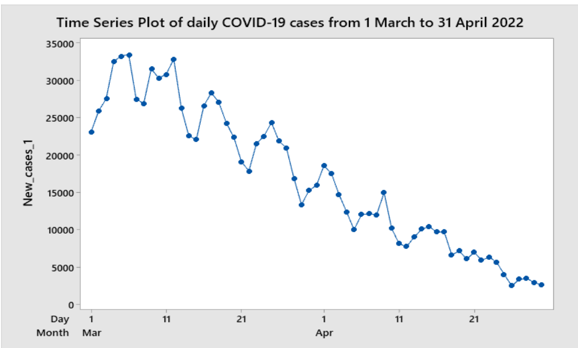

Figure 3 Time Series Plot of daily COVID-19 cases from 1 May to 31 April 2022

The time series plot above illustrates the number of daily COVID-19 cases from 1 March to 31 April 2022. It is evident that there has been a significant decrease over this period, which appears to be seasonal in nature. This indicates that there is a clear trend of decreasing COVID-19 cases during the two-month period.

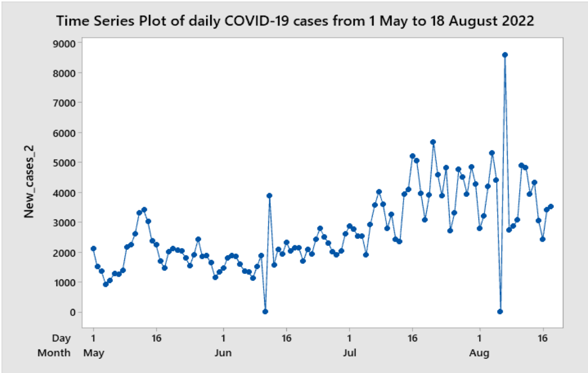

Figure 4 Time Series Plot of daily COVID-19 cases from 1 May to 18 August 2022

The time series plot above illustrates the significant increase in the number of daily COVID-19 cases over the period from 1 May to 18 August 2022. It is evident that the data exhibits a distinct seasonal pattern, indicating a potential influence of external factors in the spread of the virus. For example, the Malaysian government has officially lifted the movement control order and a number of measures such as opening up universities so that students can return to university and no longer need to use the MySejahtera mobile application to track the movements.

SARIMA

The SARIMA model is an enhanced version of the ARIMA model, designed for predicting data with both seasonal and trending patterns, whereas ARIMA is more suited for forecasting stationary time series data that includes a trend. A three-step process can be used to estimate the SARIMA model. Firstly, model identification is necessary, followed by parameter estimation, and finally, assessing model adequacy or running diagnostics.

In order to create a SARIMA model, it is essential to convert a non-stationary time series into a stationary through differencing.

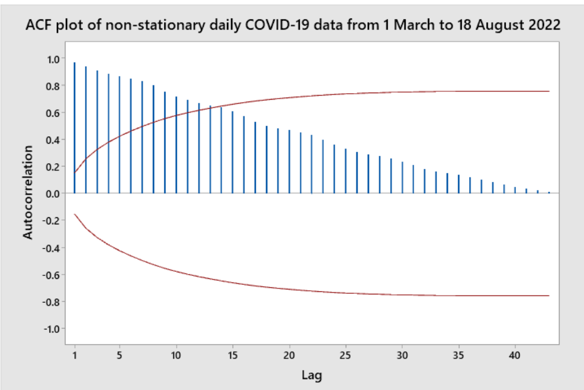

Figure 5a ACF plot of non-stationary daily 19

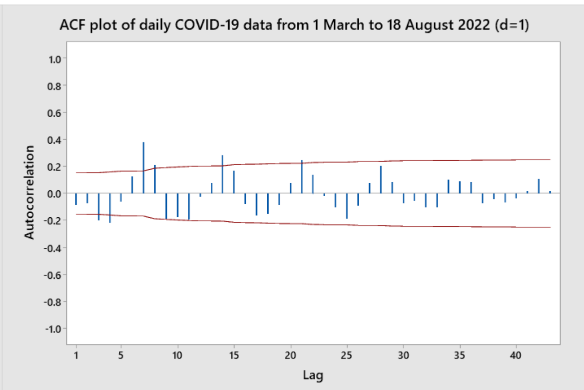

Figure 5b ACF plot of daily COVID-19 COVID- data cases with d=1

Based on the Figure 5a, the Autocorrelation Function (ACF) plot indicates that the series is non-stationary, as it decays exponentially. Hence, in order to make the non-stationary data stationary, the next step is to plot the Autocorrelation Function (ACF) again for the data. This will allow us to determine the values of the difference parameter (d) and the seasonal difference parameter (D) that will make the data stationary. From Figure 5b, there is evident that the ACF declines steadily up to 7 and then again at multiples of 7. Therefore, by setting d=1 and D= 1 and s= 7, the initial time series can be made stationary.

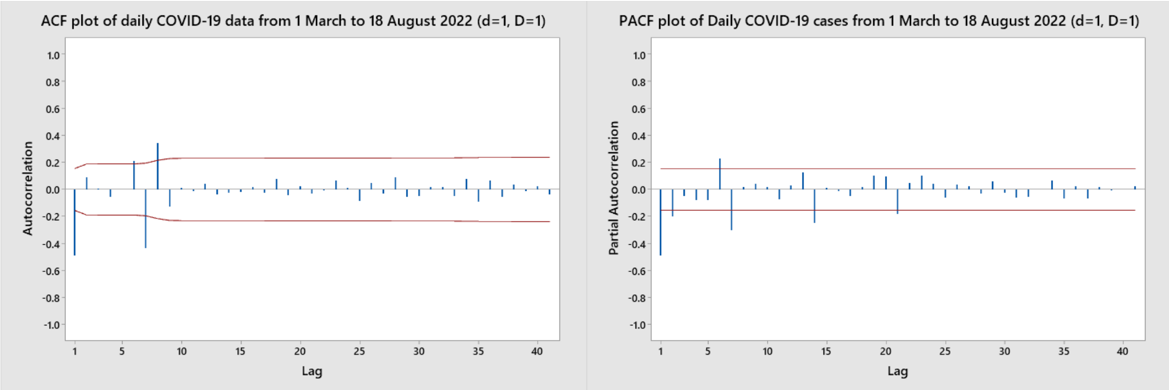

Figure 6 ACF plot and PACF plot of daily COVID-19 cases with *d*=1, *D*=1.

Stationarity of a data can be determined through visual inspection of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. ACF and PACF plots are graphical representations of the correlation between a variable and its lagged values, and can be used to identify trends and seasonality in the data. By examining the ACF and PACF plot above (Figure 6), we can discern that the majority of the peaks in both the ACF and PACF plot lie within the acceptable range of error, which suggests that the series is stationary after having been seasonally and regularly differentiated.

Once the data had been rendered stationary, we could then determine the order of the SARIMA $(p,d,q)(P,D,Q)_S$. The *d* and *D* have been filled in with 1 as both regular differencing and seasonal differencing were only conducted once. It is clear from the PACF plot that no lags beyond the second one have any significant effect. The PACF plot suggests that the most influential lags of *P* are 7, 14, and 21, implying that *P* is most likely 3. Furthermore, the ACF plot indicates that the most prominent peaks occur at lag 1. Therefore, the value of *q* is 1. We can discern the value of Q by examining the ACF plot which indicates that the peak is only significant at lag 7, thus giving us a Q of 1.

Once the data has reached stationarity, we can use SARIMA $(p,d,q)(P,D,Q)_7$ to determine the number of orders. Formerly, regular and seasonal differencing were performed once each.

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
MA 1	0.4481	0.0727	6.17	0.000
SMA 7	0.6367	0.0593	10.73	0.000
Constant	7.6	22.8	0.33	0.739

Figure 7a The Final Estimates of Parameters for SARIMA$(0,1,1)(0,1,1)_7$ and

Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

Lag	12	24	36	48
Chi-Square	23.27	28.32	35.12	40.77
DF	9	21	33	45
P-Value	0.006	0.131	0.368	0.652

Figure 7b LBQ statistics for SARIMA$(0,1,1)(0,1,1)_7$

As a result, the t statistics are significant at $\alpha = 0.05$. Moreover, the high p-values for this model indicate that the LBQ statistic is not significant. Therefore, we obtained that the model of SARIMA $(0,1,1)(0,1,1)_7$ satisfying the conditions.

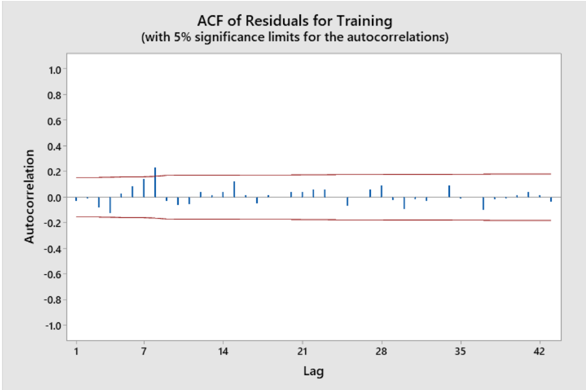

Figure 8 ACF Plot of Residual for SARIMA$(0,1,1)(0,1,1)_7$

From Figure 8, the ACF plot of residuals for SARIMA$(0,1,1)(0,1,1)_7$ are well within its standard error limit. Therefore, it indicated that the residuals are white noise. In conclusion that SARIMA$(0,1,1)(0,1,1)_7$ model best fits the data.

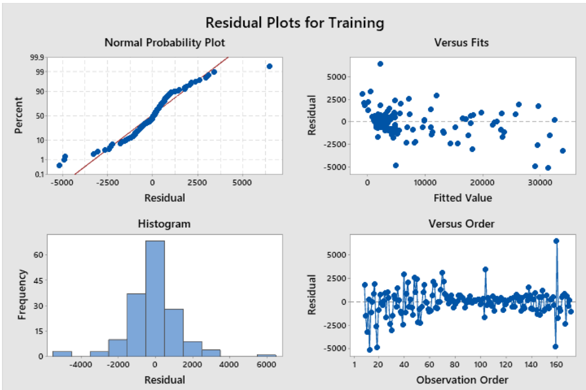

Figure 9 Residual Plots for SARIMA$(0,1,1)(0,1,1)_7$

The residuals plots showed that there are no trends or patterns, which suggests that the residuals have a relatively constant variance across the range of the data. Apart from that, the residuals are independent and normally distributed. Independence denotes that the residuals for one observation are unaffected by the residuals for other observations, which is critical for accurate predictions. The assumption of normal distribution implies that the residuals follow a bell curve and showed that the model is adequate for the data.

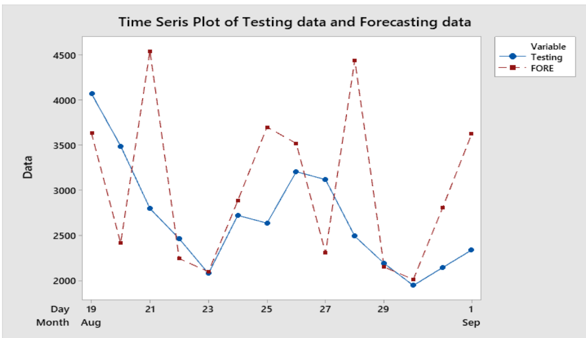

Figure 10 Time Series Plot of actual data and forecasted data.

Figure 10 eventually revealed that there is a disparity on specific spots. Especially relevant are the third, seventh, tenth, and fourteenth points.

Model Evaluation

From the result, the measurement errors of SARIMA$(0,1,1)(0,1,1)_7$ are 935.9957(RMSE), 703.8112(MAE) as well as 26.0994(MAPE).

Holt Winter Exponential Smoothing

The Additive Holt Winters method is a forecasting technique used to predict future events. This method uses an additive decomposition of the data into level, trend, and seasonal components.

The Solver function in Microsoft Excel was used to determine the most suitable values of the smoothing constants α, β, and γ that would reduce the sum of square errors to a minimum. The corresponding values of the smoothing parameters for α, β, and γ are detailed in the Table 1 below:

Table 1: Smoothing parameters using ***Solver*** function in Microsoft Excel

Smoothing parameters	Values
α	0.0267
β	0.7191
γ	0.1262

In the preceding section, the Additive Holt Winters approach has been illustrated. We will now assess the precision of the forecast by computing the MAPE, RMSE, and MAE of the out-sample data which span from 19 August to 1 September 2022.

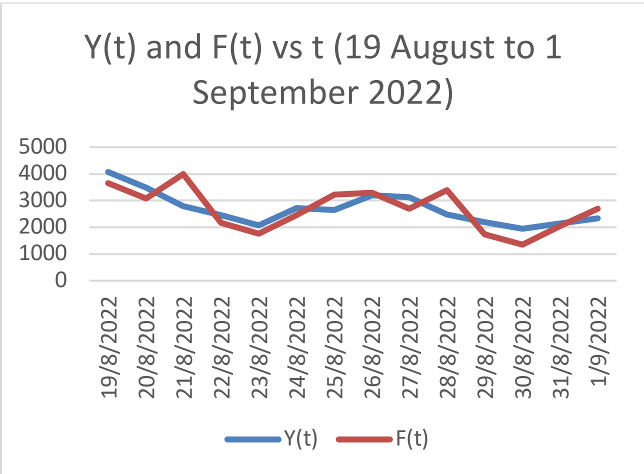

Figure 11 Actual data, Y(t) and Forecasted data, F(t) against t (19 August to 1 September)

The graph illustrates that the majority of the actual and forecasted data points are quite comparable. Despite this, a noticeable contrast can be observed in the third point, which is on 21 August 2022. Hence, the measurement errors of Addictive Holt Winter are 540.6196(RMSE), 85.4347(MAE) as well as 0.8898(MAPE).

LSTM Long Short-Term Memory (LSTM) Model

LSTM Long Short-Term Memory (LSTM) is a type of artificial recurrent neural network (RNN) architecture. It is well-suited to learning from experience to classify, process, and predict time series data when there are long-term dependencies.

In analyzing and forecasting section, the experiment is performed to analyze and predict the number of daily COVID-19 cases from 19 August to 1 September 2022 by using LSTM model.

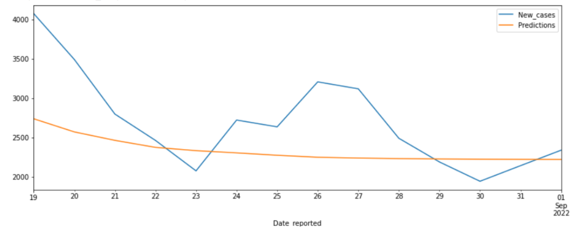

Figure 4.12 The number of actual daily COVID-19 cases and the number of predicted daily COVID-19 cases vs time, *t.*

From Figure 12, the actual data consists of a decreasing pattern and a few minor spikes. On the other hand, the predicted data appears to have a consistent decrease, without any spikes or fluctuations. This implies that the prediction model anticipates that the data will drop consistently with no variations or unexpected changes.

Afterwards, for accessing to accuracy the model the values of RMSE, MAE as well as MAPE. Therefore, the measurement errors of LSTM Model are 595.0422(RMSE), 450.7839(MAE) as well as 0.1493(MAPE).

Comparison between forecasting models

Table 2: Comparison among the forecasting models

Types of forecasting models	SARIMA	Holt Winter Exponential Smoothing	LSTM
RSME	935.9957	**540.6196**	595.0422
MAE	703.8112	**85.4347**	450.7839
MAPE	26.0994	0.8898	**0.1493**

From Table 2, there were showed the measurement error for each prediction method. By observation, the Holt Winter exponential smoothing method obtained the lowest RMSE and MAE values of 540.6196 and 85.4347 respectively. Furthermore, the LSTM obtained the lowest MAPE value of 0.1493. However, SARIMA obtained the largest values of RMSE, MAE as well as MAPE which are 935.9957, 703.8112 and 26.0994 respectively. Hence, there is obvious that the best forecasting model is Holt Winter Exponential Smoothing, then followed by LSTM and lastly SARIMA.

Conclusion

The aimed to determine the best forecasting model for daily COVID-19 cases for the next 14 days (19 August to 1 September 2022) by comparing appropriate forecasting methods based on performance or measurement error. Minitab was used to analyze trends in the daily COVID-19 case dataset and the trend analysis was divided into two parts: from 1 March to 31 April 2022 and from 1 May to 18 August 2022. The results showed a clear downward trend in COVID-19 cases over the two-month period (1 May to 31 April 2022) but a clear seasonal pattern from 1 May to 18 August 2022, indicating external factors such as the reopening of universities could have an impact on the virus spread. Forecasting models such as SARIMA, exponential smoothing and LSTM were used to predict daily COVID-19 cases for the next 14 days and their measurement error values were analyzed. The Holt Winter exponential smoothing method was determined to be the best forecasting model followed by the LSTM and then SARIMA.

References

[1] Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. Journal of Infection and Public Health, 13(7), 914–919. https://doi.org/10.1016/j.jiph.2020.06.001

[2] Aziz, N. A., Othman, J., Lugova, H., & Suleiman, A. (2020). Malaysia's approach in handling COVID-19 onslaught: Report on the Movement Control Order (MCO) and targeted screening to reduce community infection rate and impact on public health and economy. Journal of Infection and Public Health, 13(12). https://doi.org/10.1016/j.jiph.2020.08.007

[3] Bayyurt, L., & Bayyurt, B. (2020). Forecasting of COVID-19 Cases and Deaths Using ARIMA Models. https://doi.org/10.1101/2020.04.17.20069237

[4] Chudasama, H. (2020). Forecasting the Novel Coronavirus(COVID-19) using Time Series Model. Norma.ncirl.ie. https://norma.ncirl.ie/4391/

[5] Dansana, D., Kumar, R., Das Adhikari, J., Mohapatra, M., Sharma, R., Priyadarshini, I., & Le, D.-N. (2020). Global Forecasting Confirmed and Fatal Cases of COVID-19 Outbreak Using Autoregressive Integrated Moving Average Model. Frontiers in Public Health, 8. https://doi.org/10.3389/fpubh.2020.580327

[6] Ganasegeran, K., Ch'ng, A. S. H., & Looi, I. (2020). COVID-19 in Malaysia: Crucial measures in critical times. Journal of Global Health, 10(2). https://doi.org/10.7189/jogh.10.020333

[7] Gupta, R., & Pal, S. K. (2020). Trend Analysis and Forecasting of COVID-19 outbreak in India. https://doi.org/10.1101/2020.03.26.20044511

[8] Hebbar, N. (2022, August 6). Time-Series-Forecasting-LSTM. GitHub. https://github.com/nachi-hebbar/Time-Series-Forecasting-LSTM

[9] Kırbaş, İ., Sözen, A., Tuncer, A. D., & Kazancıoğlu, F. Ş. (2020). Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos, Solitons & Fractals, 138, 110015. https://doi.org/10.1016/j.chaos.2020.110015

[10] Lone, S. A., & Ahmad, A. (2020). COVID-19 pandemic – an African perspective. Emerging Microbes & Infections, 9(1), 1300–1308. https://doi.org/10.1080/22221751.2020.1775132

[11] Rahimi, I., Chen, F., & Gandomi, A. H. (2021). A review on COVID-19 forecasting models. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05626-8

[12] Sarkar, S. (2020, April 12). TIME SERIES FORECASTING AND ANALYSIS : ARIMA AND SEASONAL-ARIMA. Analytics Vidhya. https://medium.com/analytics-vidhya/time-series-forecasting-and-analysis-arima-and-seasonal-arima-cacaf61ae863

[13] Syafiq, M., Alza, A., Norazlin, I., & Zalhan, M. Z. (2021). Comparison of Performance for Forecasting the New Cases of Covid-19 In Malaysia by Using ARIMA, LSTM and Prophet. https://doi.org/10.17605/OSF.IO/NM3K8

[14] Tan, C. V., Singh, S., Lai, C. H., Zamri, A. S. S. M., Dass, S. C., Aris, T. B., Ibrahim, H. M., & Gill, B. S. (2022). Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia. International Journal of Environmental Research and Public Health, 19(3), 1504. https://doi.org/10.3390/ijerph19031504

[15] Valkov, V. (2023, May 8). Get SH*T Done with PyTorch. GitHub. https://github.com/curiousily/Getting-Things-Done-with-Pytorch

[16] Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. Chaos, Solitons & Fractals, 140, 110214. https://doi.org/10.1016/j.chaos.2020.110214

[17] World Health Organization. (2021a). The true death toll of COVID-19: estimating global excess mortality. Www.who.int. https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality

[18] World Health Organization. (2021b). Who.int. https://covid19.who.int/data

[19] World Health Organization. (2022). Weekly epidemiological update on COVID-19 - 3 August

2022. Www.who.int. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---3-august-2022

[20] Yurinanda, S., Multahadah, C., & Aryani, R. (2020). Development of COVID-19 Case in District and City of Jambi Province with Exponential Smoothing Methode. EKSAKTA: Berkala Ilmiah Bidang MIPA, 21(2), 110–123. https://doi.org/10.24036/eksakta/vol21-iss2/244