INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 11, Issue 6, (2021) E-ISSN: 2222-6990

Forecasting Dengue Outbreak Data Using ARIMA Model

Nur Syuhada Muhammat Pazil[1], Norwaziah Mahmud[2], Siti Hafawati Jamaluddin[3] and Nur Aqilah Ali[4]

[1]Universiti Teknologi MARA Cawangan Melaka Kampus Jasin, Melaka MalaysiaJasin Melaka, Malaysia, [2,3,4]Universiti Teknologi MARA Cawangan Perlis Kampus Arau, Perlis, Malaysia, 02600 Arau Perlis, Malaysia

Email: syuhada467@uitm.edu.my, norwaziah@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v11-i6/10106 DOI:10.6007/IJARBSS/v11-i6/10106

Published Date: 21 June 2021

Abstract

Dengue fever is an internationally recognised virus that is spread by mosquitoes and can result in death. From recorded cases, Selangor has the highest rate of dengue infection among Malaysian populations. Corona disease (COVID-19), a new pandemic that has swept the globe, like Selangor, has prompted this report on the pattern of dengue cases during COVID-19 pandemic. Due to the new outbreak of COVID-19, the Movement Control Order (MCO) has been extended from time to time, and with most health resources at the state and federal levels being used to combat COVID-19, dengue control activities have been limited to a non-contact activity in outbreaks and hotspot areas. The importance of this study is to investigate the increase in dengue cases in Selangor because Selangor recorded the highest number of cases in Malaysia. Considering that there are not many studies conducted in Selangor, this study is important to predict dengue cases, and the authorities can take immediate action to overcome this problem. The aim of this research is to find the best ARIMA model for predicting the dengue cases in Selangor in the future. Several ARIMA models were used to test dengue cases data obtained in Selangor in order to achieve the objectives. The best model was calculated by comparing the Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percent Error (MAPE) measurement errors. Then, the predicted number of dengue cases was calculated using the best model generated. The model for which the values of criteria are the smallest is considered as the best model. Hence ARIMA (1,1,2) was found to be the best model for predicting the dengue cases data series and this model is used to predict the number of dengue cases in Selangor with the smallest Mean Square Error (MSE) value of 12837.4327 and Root Mean Square Error (RMSE) value of 113.3024. The forecasted values showed a decreasing number of dengue cases. This study was carried out using R-studio software and excel. Further research can be conducted using another time series method, for example Holt-winters method.

Keywords: Prediction, Dengue, MSE, ARIMA Model, RMSE, MAPE

Introduction

Dengue is one of the most serious and fast emerging tropical viruses that is transmitted by female mosquitoes mainly of the species *Aedes aegypti* and, to a lesser extent, *Ae. Albopictus*. Mosquitoes can thrive in places with standing water, such as water tanks, buckets, and stagnant pools of water. According to the World Health Organization (WHO), about 2.5 billion people are currently living in dengue-infested areas. There are four dengue serotypes that cause dengue outbreaks designated as DEN-1, DEN-2, DEN-3, and DEN-4 (Martinez et al., 2011). The disease is endemic in 128 countries throughout the South Asia, South East-Asia, Africa, the Americas, the Western Pacific and Eastern Mediterranean Region. (Salim et al., 2021)

From the European Centre for Disease Prevention and Control (2020), the Pan American Health Organisation (PAHO) reported statistics in the Americas region, showing the highest dengue cases recorded in Brazil with 167,000 cases, followed by Paraguay with 85,000 cases and 20,000 dengue cases recorded in Colombia on 9 February 2020.

Malaysia is also one of the countries with high number of dengue cases among the states. However, Selangor ranked the highest number of dengue cases every week of the year. Although several types of research have been conducted, there are still a large number of dengue cases in Selangor. In a previous study, Mudin (2015) reported that there are a number of epidemic peak incidence factors in Malaysia. Other than changes in serotypes, climate change factors, such as increased rainfall, high temperatures and increased air humidity, have led to a high chance of dengue transmission and its occurrence in Malaysia. As Malaysia's economy and industry grow rapidly, this is also becoming a factor in the spread of viruses. Human movement has caused dengue to spread widely to other places as they go there. The dengue virus can be infected to humans at any stage of age. Normally, those infected with dengue virus will have symptoms such as headache, rash, vomiting, nausea, swollen glands, muscle and joint pain, and may feel pain behind their eyes (World Health Organisation, 2020).

Several studies have been carried out to predict dengue outbreaks in Selangor, which has the highest number of dengue cases in Malaysia, by using time series analysis. According to Dom et al (2013), forecasting model for the dengue cases in Subang Jaya was performed using the Autoregressive Integrated Moving Average (ARIMA) and it concluded that the ARIMA model with weekly variation is an important tool for disease control and prevention programmes in Malaysia because it can accurately predict the number of dengue cases. Shah and Sani (2011) indicated that dengue fever incidence can be predicted using time series analysis, which can help with long-term preparation of dengue fever management and prevention programme.

ARIMA stands for the Autoregressive Integrated Moving Average model, which is one of the methods used in the time series analysis for forecasting. There are two types of ARIMA method; the non-seasonal Autoregressive Integrated Moving Average and the seasonal Autoregressive Integrated Moving Average (SARIMA). However, in this study, non-seasonal ARIMA will be used to predict dengue outbreaks as it is a time-series analysis and this method is one of the best fit models used in previous studies. Sato (2013) mentioned that George Box and Gwilym Jenkins had designed the ARIMA model in the 1970s to define time series deviations using a mathematical method. ARIMA and Box-Jenkins names are declared synonyms. The objective of this model is to modify the observed values and to minimise the difference between the values obtained in the model and the observed values as close to zero as possible.

Fattah et al (2018) used ARIMA models to predict the demand in a food company. By using Box-Jenkins time series approach, they developed an ARIMA model for forecasting. The researchers found that, ARIMA (1,0,1) shows better performance according to four performance criteria which are SBC, AIC, standard error and maximum likelihood.

Paul and Hoque (2013) used ARIMA model in forecasting the average daily share price (ADSPI) of the Square Pharmaceuticals Limited (SPL). The study found that ten tentatively ARIMA models by using computer software SHAZAM version 8, ARIMA (1,1,1), ARIMA (1,1,2), ARIMA (2,1,1), ARIMA (2,1,2), ARIMA (1,1,3), ARIMA (2,1,3), ARIMA (3,1,1), ARIMA (3,1,2), ARIMA (3,1,3) and ARIMA (1,1,4). Then, the best model was determined by using the selected criteria AIC, AICc, SIC, AME, RMSE, and MAPE in the three periods which were the estimation period, validation period and total period. Hence, ARIMA (2,1,2) model has been selected for forecasting by referring the maximum number of the lowest values of all the selected criteria.

Singh et al (2020) forecasted daily confirmed COVID-19 cases in Malaysia using ARIMA model. The study showed that ARIMA (1,0,1) model fits the daily COVID-19 cases in Malaysia satisfactorily. Erlina and Azhar (2020) used ARIMA in predicting the monthly series of export coffee from January 2005 to April 2020. The results show that ARIMA (1,3,1) is the best selected model due to its significant p-value. Flood prediction for Pengkalan Rama, Melaka river using ARIMA model was studied by Wong et al. (2020). The best ARIMA model was identified by the parameter Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The best ARIMA model for the Pengkalan Rama is ARIMA (2, 1, 2).

From the previous studies, it is clear that ARIMA can be used to forecast. Hence, this study aims on selecting the best ARIMA model and to forecast dengue cases in Selangor. The data used for this analysis are secondary data provided by the Vector Borne Disease Control Section of the Ministry of Health Selangor. The study focuses on the collection of data in Selangor. The data showed epidemic weekly dengue outbreak data from January 2018 to November 2020. To select the best ARIMA model, their goodness of fit had been compared using the Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percent Error (MAPE) measurement errors by looking at their smallest error values.

Methodology

The ARIMA model was generated in this study to predict the outbreak of dengue. This model has been identified as one of the most effective methods used to predict time-series data. This method is divided into three phases, which are identification, estimation and forecasting phases. The actual data have to be stationary. If the data is not stationary, proceed to the identification stage.

Identification Phase:

Step 1: Evaluate the stationary data series with three approaches.

i) Plot time series plot for the actual data and check if a trend component appears on the plot.

ii) Plot the autocorrelation function (ACF) correlogram and partial autocorrelation function (PACF) correlogram. If decay occurs in the ACF plot, the stationary ACF plot can be determined by applying the Augmented Dicky-Fuller (ADF) test.

Decision Rule:
Reject H_0 if p-value ≥ 0.05
H_0: The data is stationary.
H_1: The data is not stationary.

iii) Check the PACF plot if there is a higher spike shown in the plot then the data is not stationary.

Step 2: Differencing when the data is non-stationary.

Differencing can also be checked using the Kwiatkowski–Phillips–Schmidt–Shin test (KPSS Test) whether or not differentiation is required. If the test statistics shown are greater than 10% of the critical value, the null hypothesis is rejected. Consequently, the data is not stationary. Repeat the data differencing and use the KPSS test again.

Decision Rule:
Reject H_0 if p-value > 0.1
H_0: The data is stationary.
H_1: The data is not stationary.

Estimation Phase:

Step 3: Check the parameters estimated for the ARIMA model.

The ARIMA model is categorized into three terms, p, d, and q, where p is the order of the autoregressive (AR) term, q is the order of the moving average (MA) term, and d is the number of differencing used to obtain the stationary time series data. These are the mathematical formulas for each term:

Autoregressive (AR), p term,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \qquad (1)$$

Moving Average (MA), q term,

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \ (2)$$

Where $\varepsilon_t$represents white noise.

Number of differencing, d term,

$$\triangle^k y_t = (1-B)^k y^t \ (3)$$

Where B is the lag operator.

Therefore, the general formula for the ARIMA model is as follows,

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

Where $y'_t$is a differenced series.

Forecast Phase:
Step 4: Check the accuracy of the model to select the best ARIMA model. To verify the accuracy of the model, measuring errors of the ARIMA model, such as MSE, RMSE, and MAPE, are then calculated. Then, the best model is selected to forecast the data.

Findings and Discussions
By referring to the pattern of the actual data of dengue cases in Figure 1, it shows that there was a current trend in the dengue data in Selangor between January 2018 and November 2020. Since the data show the presence of a trend in the graph, the data are not stationary and proceed to the identification phase.

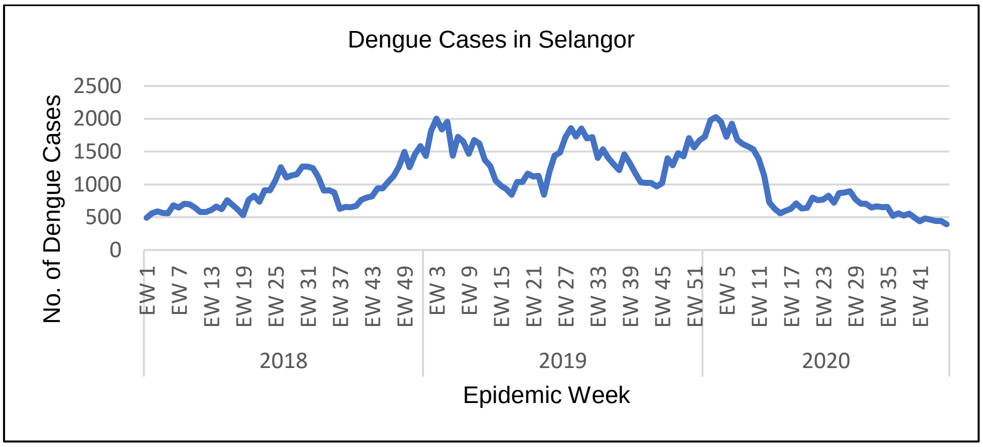
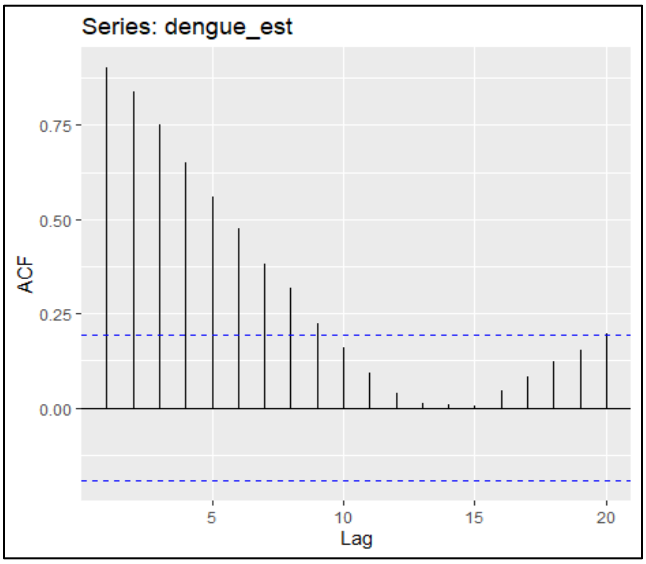
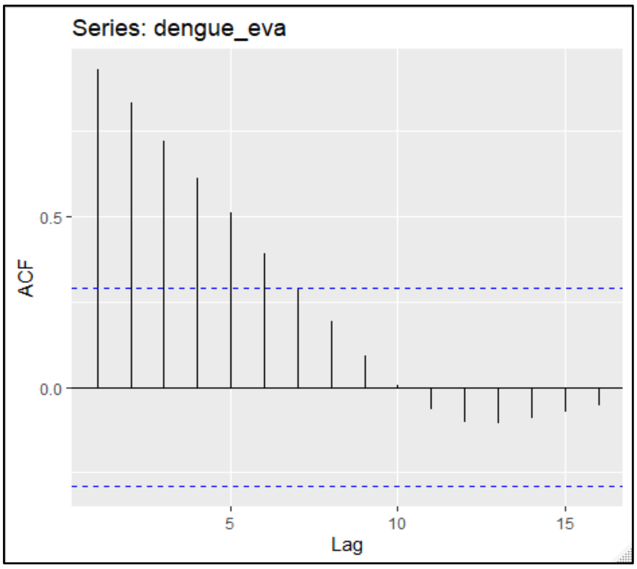


Figure 1: Weekly dengue cases in Selangor plot

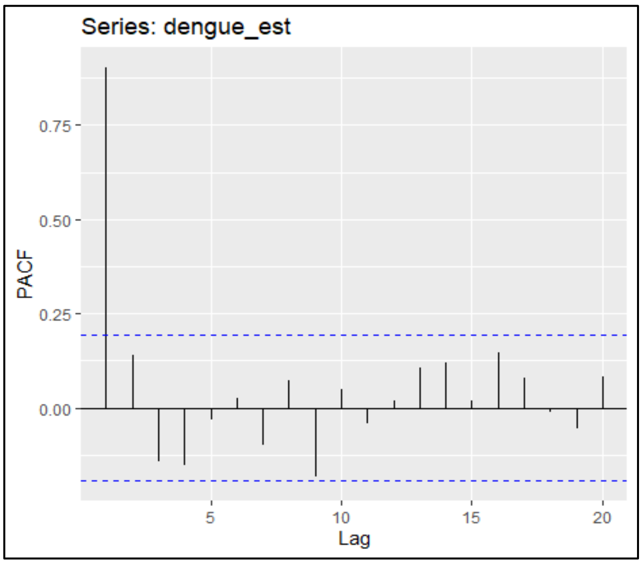
The data were divided into two parts which are estimation (70%) and evaluation (30%) parts. In the estimation part, the data selected are from EW1 2018 (n=1) to EW52 2019 (n=104). Data from EW1 2020 (n=105) to EW46 2020 (n=150) were used in the evaluation part.
The original data were evaluated using three approaches; the ACF plot, the PACF plot and the ADF test for the estimation part and evaluation part.





ACF Plot

PACF Plot



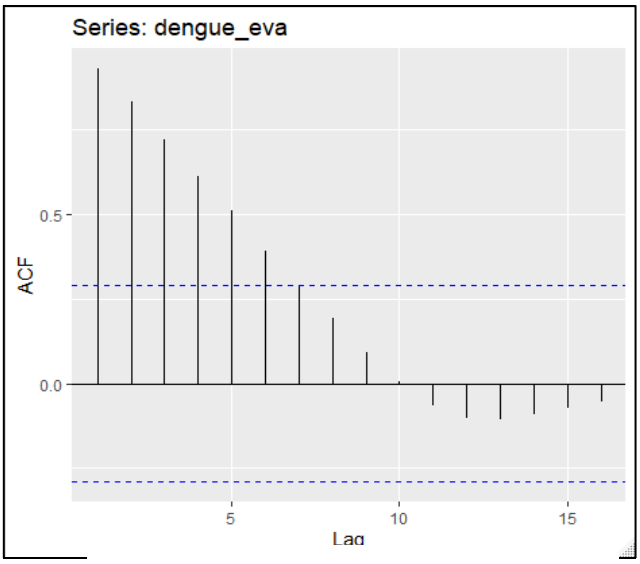


Figure 4: PACF plot of estimation part

Figure 5: PACF plot of evaluation **part**

The pattern of decay is shown from the ACF plot in Figure 2 and Figure 4. Figure 3 and Figure 5 show the PACF plot of the estimation part and the evaluation part respectively, with a large spike. Therefore, the data are not stationary.

Next, the ADF test is generated by using the R-studio software where the p-value is computed for the estimation and evaluation parts.

Table 1
Output of ADF test for estimation part

Statistics	**Model**
Decision Rule: Reject H_0 if p-value ≥ 0.05.	*H_0*: The data is stationary. *H_1*: The data is not stationary.
P-value	0.0808
Decision (5% significant level)	Reject *H_0*
Conclusion	The data is not stationary.

Table 2
Output of ADF test for evaluation part

Statistics	**Model**
Decision Rule: Reject H_0 if p-value ≥ 0.05.	*H_0*: The data is stationary. *H_1*: The data is not stationary.
P-value	0.4289
Decision (5% significant level)	Reject *H_0*
Conclusion	The data is not stationary.

Since all the alternatives show that the data of the estimation and evaluation parts are not stationary, thus the first differencing should be done by checking the KPSS test.

Table 3
Output of KPSS test

Statistics	**Model**
Decision Rule: Reject *H_0* if p-value > 0.1	*H_0*: The data is stationary. *H_1*: The data is not stationary.
P-value	0.043166
Decision (10% significant level)	Accept *H_0*
Conclusion	The data is stationary.

After verification of the stationary on the first differencing, several models were tested to identify the most suitable one. Ten ARIMA models with tentatively selected values p, d, q were estimated and among the models, only five were found to be comparatively well performed models (Paul & Hoque, 2013). Hence, the measurement errors of the five ARIMA models, which are MSE, RMSE and MAPE, were computed.

Table 4
Measurement errors of the five ARIMA models

Measurement Errors	**ARIMA (1,1,1)**	**ARIMA (1,1,2)**	**ARIMA (1,1,3)**	**ARIMA (2,1,1)**	**ARIMA (2,1,2)**
MSE	13396.399	12837.4327	13076.82082	12851.78961	13166.73706
RMSE	115.742814	113.3023949	114.3539	113.3657	114.74640
MAPE	9.2467	8.82887	8.81120	8.9948	8.7796

Table 4 shows that ARIMA with p=1, d=1, q=2 process maximum number of the lowest measurement errors. Hence, ARIMA (1,1,2) model has been selected as the most suited model for forecasting the number of dengue cases in Selangor. The predicted number of dengue cases were computed using the best model selected which is ARIMA (1,1,2) model as shown in Table 5.

Table 5
Predicted number of dengue cases in Selangor

	Predicted Dengue Cases
Epidemic Week 47	385
Epidemic Week 48	372
Epidemic Week 49	369
Epidemic Week 50	368
Epidemic Week 51	368
Epidemic Week 52	367

From Table 5, the number of dengue cases predicted in Selangor on Epidemic Week 47 until Epidemic Week 52 kept on decreasing from 385 cases down to 367 cases.

Conclusion and Recommendations

The best ARIMA model~~s~~ to predict dengue case values in Selangor for Epidemic Week 47 until Epidemic Week 52 in 2020 can be generated by conducting this study. As a result, the best model was evaluated by the least measurement errors of the models used in the analysis. The method chosen for the study should provide the results of the prediction that reflect the actual behaviour of the time series data. This study was therefore analysed using ARIMA (1,1,2), since the model was chosen as the best method for assessing the actual dengue case data in Selangor from Epidemic Week 1 January 2018 to Epidemic Week 46 November 2020.

From the actual data, the highest number of dengue cases recorded in Epidemic Week 3 January 2020 is 2026 while the lowest number of reported dengue cases from the collected data is 392 cases in Epidemic Week 46 November 2020. Dengue cases have shown a trend where there was an increase in the number of cases in 2019. However, as the new pandemic (COVID-19) occurred in early 2020, it shows that the number of dengue cases started to drop dramatically from Epidemic Week 3 to Epidemic Week 16. The number of dengue infections did not increase as much as before the occurrence of COVID-19. After EW16, the number of dengue cases dropped to under 1000 cases and continued to decline.

The analysis shows the number of dengue cases during the new pandemic, with COVID-19 still occurring among the community in Selangor in the year 2020. In order to accurately predict future dengue case values that may occur in Selangor for the weeks to come, ARIMA (1,1,2) was assessed as the best model to be used. With the predicted values produced, the study showed a pattern in which dengue cases decreased during COVID-19, similar to the actual dengue case data collected. There are so many variables that have caused this to happen during COVID-19 where the number of dengue cases gets lower. One of the factors is that the community is less involved with each other when staying at home, so the cases of infection are reduced. The analysis on this topic recommended the study to be conducted using more time series methods to evaluate the accurate predicted dengue case values. With the smallest measurement errors computed by each method, the model generated is the best model to provide high accuracy of the predicted dengue cases.

References

Dom, N. C., Hassan, A. A., Latif, Z. A., & Ismail, R. (2013). Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia. *Asian Pacific Journal of Tropical Disease*, *3*(5), 352–361. https://doi.org/10.1016/S2222-1808(13)60084-5

Erlina, R., & Azhar, R. (2020). Forecasting model of agriculture commodity of value export of coffee; Application of Arima Model. *Jurnal Teknik Pertanian Lampung (Journal of Agricultural Engineering)*, *9*(3), 257. https://doi.org/10.23960/jtep-l.v9i3.257-263

European Centre for Disease Prevention and Control (2020). *Dengue worldwide overview*. Retrieved from https://www.ecdc.europa.eu/en/dengue-monthly

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, *10*, 1–9. https://doi.org/10.1177/1847979018808673

Martinez, E. Z., Silva, E. A. S. Da, & Fabbro, A. L. D. (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Revista Da Sociedade Brasileira de Medicina Tropical*, *44*(4), 436–440. https://doi.org/10.1590/s0037-86822011000400007

Mudin, N. R. (2015). Dengue incidence and the prevention and control program in Malaysia. *International Medical Journal Malaysia*, *14*(1), 5–9. https://doi.org/10.31436/imjm.v14i1.447

Paul, J. C., & Hoque, S. (2013). *Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd. 13*(3).

Salim, N. A. M., Wah, Y. B., Reeves, C., Smith, M., Yaacob, W. F. W., Mudin, R. N., Dapari, R., Sapri, N. N. F. F., & Haque, U. (2021). Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. *Scientific Reports*, *11*(1), 1–9. https://doi.org/10.1038/s41598-020-79193-2

Sato, R. C. Esa. (2013). Disease management with ARIMA model in time series. *Einstein (São Paulo, Brazil)*, *11*(1), 128–131. https://doi.org/10.1590/S1679-45082013000100024

Shah, S. A., & Sani, J. A. (2011). SP6-37 predicting dengue fever incidence in Selangor using time series analysis technique. *Journal of Epidemiology & Community Health*, *65*(Suppl

1), A464–A464. https://doi.org/10.1136/jech.2011.142976q.8

Singh, S., Sundram, B. M., Rajendran, K., Law, K. B., Aris, T., Ibrahim, H., Dass, S. C., & Gill, B. S. (2020). Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. *Journal of Infection in Developing Countries*, *14*(9), 971–976. https://doi.org/10.3855/JIDC.13116

Wong, W. M., Subramaniam, S. K., Feroz, F. S., Subramaniam, I. D., & Rose, L. A. F. (2020). Flood prediction using ARIMA model in Sungai Melaka, Malaysia. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5287–5295. https://doi.org/10.30534/IJATCSE/2020/160942020

World Health Organization (2020). Dengue and severe dengue. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue