INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 14, Issue 1, (2024) E-ISSN: 2222-6990

A Comparison between ARIMA and Fuzzy Time Series Methods in Predicting Daily COVID-19 Outbreak in Malaysia

Nur Syuhada Muhammat Pazil[1], Siti Nor Nadrah Muhamad[2] and Norsuhaila Sulaiman[3]

[1]Mathematical Sciences Study, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Melaka Branch, Jasin Campus, 77300 Merlimau, Melaka, Malaysia.[2,3]Mathematical Sciences Study, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau,Perlis, Malaysia.

Corresponding Auhtor Email: nadrahmuhamad@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v14-i1/18359 DOI:10.6007/IJARBSS/v14-i1/18359

Published Date: 05 January 2024

Abstract

COVID-19 is a viral infection caused by a recently identified coronavirus that has impacted the lives of millions of people worldwide. In Malaysia, the number of COVID-19 cases has been increasing since 2021. This study aims to find the best model for forecasting the number of new confirmed cases of COVID-19 in Malaysia by comparing Autoregressive Integrated Moving Average (ARIMA) and Fuzzy Time Series models. ARIMA is commonly used for time-series analysis, forecasting, and control, while Fuzzy Time Series provides an alternative approach for predicting COVID-19 outbreaks. The error measures used to compare the models include Mean Square Error, Root Mean Square Error, and Mean Absolute Percentage Error. The study's results demonstrate that the Fuzzy Time Series model has the smallest error measure values compared to ARIMA, indicating that it is more accurate.

Keywords: ARIMA, COVID-19, Fuzzy Time Series, Mean Square Error; Error Measure

Introduction

In early 2020, the world was shocked by the emergence of COVID-19, a deadly infectious disease that affected millions of people globally. No one could have imagined that a virus or disease could have such a profound impact on the world. The World Health Organization (WHO) defines COVID-19 as an infectious disease caused by a newly discovered coronavirus (WHO, 2020). Common symptoms of COVID-19 include fever, dry cough, body aches, nasal congestion, runny nose, sore throat, diarrhea, and vomiting (D. Wang et al., 2020).

The COVID-19 outbreak was first reported in late December 2019 in a seafood wholesale wet market, the Huanan Seafood Wholesale Market, in Wuhan, Hubei, China (Huang et al., 2020). The disease spread rapidly among the residents of Wuhan City and later to other countries such as Japan, the Republic of Korea, Thailand, Viet Nam, the United States, Germany and Singapore (Wu et al., 2020). The first COVID-19 case outside China was reported in Thailand on January 13, 2020 (WHO, 2020).

On January 25, 2020, Malaysia reported its first COVID-19 positive case and which was an imported case from Wuhan, China. The first Malaysian test positive for COVID-19 was reported on February 3, 2020 (Bernama, 2022). In March 2020, the situation became worse as there was a spike in cases resulting from the religious gathering cluster in Sri Petaling, Selangor (Babulal & Othman, 2020). In response to the situation, the government has imposed a Movement Control Order to control the spread of the COVID-19 infection in Malaysia. After about three to five months, the numbers of COVID-19 infections in Malaysia decreased. Unfortunately, in 2021, the number of COVID-19 cases began to increase again and continue to rise.

Autoregressive Integrated Moving Average (ARIMA) modelling is commonly applied to the time-series analysis, forecasting and control. In the study by Pazil, N. S. et at. (2021) the ARIMA model was used to predict the dengue outbreak in Selangor. Several ARIMA models were tested to determine the best fit by comparing the value of goodness of fit. Mahmud et al. (2021) used the ARIMA models to forecastdengue outbreak and compared them with other models. According to Singh et al. (2020), ARIMA models with carefully chosen covariates are effective tools for tracking and forecasting trends in COVID-19 cases in Malaysia. The result of the study showed that the ARIMA (0,1,0) model produced the best fit to the observed data with a Mean Absolute Percentage Error (MAPE) value of 16.01 and a Bayes Information Criteria (BIC) value of 4.170. The predicted values indicated a decline in COVID-19 cases through May 1, 2020.

A study on forecasting the dynamics of the COVID-19 epidemic in India using ARIMA models found that it is easy to apply and interpret at both national and regional levels to monitor the spread of COVID-19 (Katoch & Sidhu, 2021). Another study on forecasting COVID-19 confirmed cases in Korea using empirical data analysis found the ARIMA model has relatively excellent and stable predictive power (Lee et al., 2021). The purpose of the study was to emphasize on the importance of prediction timing rather than prediction of the number of confirmed cases. The results showed that the ARIMA model was well-fitted and effectively estimate the number of confirmed cases.

According to Mishra et al. (2020), in their study about the trajectory of COVID-19 data in India using Artificial Neural Network, Fuzzy Time Series, and ARIMA models, they found out that the ARIMA model is more appropriate to forecast the virus trajectories. The results showed that the root mean square error (RMSE) value of ARIMA was smaller for both new cases and new deaths time series, with values of 436.55 and 31.08, respectively. The result indicated that the ARIMA model fits the data better compared to Fuzzy Time Series model.

In addition, a study on forecasting the COVID-19 outbreak using ARIMA and Fuzzy Time Series models stated that the ARIMA model gives remarkably satisfying results in predicting natural adversities compared to other predictions models (Verma et al., 2020). The data used in this study was COVID-19 new cases, infections, mortalities, and recoveries obtained from the Ministry of Health and Family Welfare, Government of India. The results showed that both, ARIMA and Fuzzy models suggest an increasing number in COVID-19 cases in the future. The

study found that the Fuzzy Time Series model can be an alternative way to forecast COVID-19 sepsis.

Fuzzy Time Series is a new concept proposed by the authors to deal with forecasting problems where the historical data consists of linguistic values. The concept of the fuzzy time series was first developed by Song and Chisson, as noted by Verma et al. (2020). The main advantage of the Fuzzy Time Series technique is that no assumptions are made regarding the data set.

According to Cai et al. (2013), Fuzzy Time Series is an application of fuzzy mathematics in the field of time series. If the data is complete and contains noise, using fuzzy theory to help forecast can generally obtain better results. The dataset used in this study was the TAIEX stock market. The results showed that a hybrid model FTSGA, based on fuzzy time series and genetic algorithm, improved the performance and accuracy by applying genetic algorithm to the operations.

Furthermore, Alam et al. (2022) proposed a new procedure for forecasting time series data based on the intuitionistic fuzzy set (IFS). The proposed model also utilizes a defuzzification formula that fully utilizes the main properties of IFS, which include the membership and non-membership functions. The defuzzification procedure used in this study significantly decreased forecasting error. Even with the volatility of the Crude Palm Oil (CPO) price,this model produced noteworthy forecasting results.

Another study applied the weighted fuzzy time series model to forecast epidemic injuries and found that it provides significantly better results than the classical statistical methods (Moneim, 2020). The study used COVID-19 epidemic injury data in Saudi Arabia from March 2nd, 2020 to July 20th, 2020. The results showed that the mean square error of the Weighted Fuzzy Time Series (WFTS) method was 0.0049 which was smaller than the previous methods. This effectively demonstrated that WFTS method is a good tool for predicting COVID-19 epidemic injuries.

As one of the successful countries in handling the COVID-19 pandemic globally, Malaysia ~~has~~ needs to ensure that the number of new confirmed cases of COVID-19 always decreases and that the number of COVID-19 recovered cases increases to avoid a worsening situation. However, in 2021, the number of new confirmed cases recorded in Malaysia has been increasing day by day. This may lead toa lockdown for the whole country and if the daily cases continue to rise, the government will have to implement a total lockdown, in which would involve closing the entire sector, and people would not be allowed to go outside. These problems will affect the government, especially front liners, people and the economy of Malaysia. Therefore, a precise forecasting model is needed to assist the government in predicting the future value of the new confirmed and recovered cases of COVID-19. The forecasting model will also help the government make a good plan to prevent losses in handling the COVID-19 pandemic.

Hence, this study focuses on comparison between Autoregressive Integrated Moving Average (ARIMA) and Fuzzy Time Series for predicting new cases of COVID-19 in Malaysia. These time series methods aim to accurately predict the forecasted data by minimizing error values.

Methodology

Data Collection Method

The data used for this analysis comprises the daily data on newly confirmed cases of COVID-19 in Malaysia from July 2021 until December 2021. Both Autoregressive Integrated Moving Average (ARIMA) and Fuzzy Time Series techniques were used to compare the data and

identify the best technique. RStudio was used for ARIMA while Microsoft Excel was used for Fuzzy Time Series. The techniques were compared based on their Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE) performances to determine which technique is superior.

ARIMA Model

ARIMA models are widely used techniques for time series forecasting (Verma et al., 2020). This model has been identified as one of the most effective methods to predict time-series data. The ARIMA method is divided into three phases, which are identification, estimation and forecasting phases. The actual data should be stationary. If the data is not stationary, proceed to the identification stage. Evaluate the stationary data series with three approaches, which are ACF, PACF and ADF test. Difference the data when it is still non-stationary. Next, proceed with the estimation phase by checking 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.

Autoregressive (AR), p term,

$$y_t = c + \phi_1 y_{t-1} + \phi_1 y_{t-2} ..... + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

Moving Average (MA), q term,

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

Where ε_t represents white noise.

Number of differencing, d term,

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

where B is the lag operator.

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

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} ..... + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

Where, y'_t is a differenced series.

In this study, ARIMA (2,1,8) was computed for the ARIMA model from the result generated using R-studio software. The general equation can be expressed using equation 4.

Fuzzy Time Series

Song and Chisson first developed the concept of the fuzzy time series (Verma et al., 2020). There were seven steps or algorithms presented in the Fuzzy Time Series model.

Step 1: All the data were analyzed and then changed it into a percentage form. The formula shown below is a calculation of the percentage form.

$$y_n = \frac{y_n - y_{n-1}}{y_n} \times 100 \quad (5)$$

Where, y_n is the number of new cases of COVID-19, y_{n-1} is the number of new cases of COVID-19 before.

Step 2: Define the universe, *U* : Two minimum and maximum values have been identified from the percentage form in step 1. The maximum value is 29.22% while the minimum value is -24.56%. The universe of discourse, *U* was needed by finding the minimum enrollment D_{min} and the maximum enrollment D_{max}. The formula universe of discourse *U* is defined by:

$$U = \left[D_{\min} - D_1, D_{\max} + D_2 \right] \quad (6)$$

Step 3: The fuzzy sets U_i were constructed using the same length of intervals where it starts from 1 until 7. In this step, the fuzzification of interval and frequency of each interval were identified. The formula shown below is a calculation of the length of interval for fuzzification.

$$\frac{\left(D_{\max} + D_2\right) - \left(D_{\min} + D_1\right)}{7} \quad (7)$$

Step 4: The interval of $v_1, v_2, \ldots, v_n$, were generated based on step 2 and the interval were formed in trapezoidal number.

Step 5: Each of the data was listed and classified in percentage forms and based on the interval in step 4. Thus, the fuzzy logical relationship was identified from the classified data. Fuzzy logical relation was represented as $A_i \rightarrow A_j$where A_i is presented in form and A_j is the future form.

Step 6: A fuzzy logical relationship rule is required to identify and is based on the fuzzy logical relations from step 5. The fuzzy logical relationship rule must be generated in groups.

Step 7: Each fuzzy logical relationship group needs to be classified into one of the three different types of rules. Every calculation of fuzzy logical relationship was different based on the rules below.

Rule 1: The fuzzy group of A_j is null or empty, $A_j \rightarrow \emptyset$ or same, $A_j \rightarrow A_j$.
Rule 2: The fuzzy group of A_j is a one-to-one relationship for example $A_j \rightarrow A_k$.
Rule 3: The fuzzy group of A_j is a one-to-many relationship, for example $A_j \rightarrow A_k, A_l$.

Evaluate Model Performance

From the generated models, the goodness of model in which the measurement of errors for accuracy of the forecasted data was obtained from the actual data and the ARIMA and the Fuzzy Time Series technique were compared to the lowest error measurements. There are few measurements of error that can be tested, such as Mean Squared Error (MSE), Root Mean

Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These error measurements were formulated as follows:

$$\text{Mean Squared Error (MSE)} = \frac{\sum e_t^{\,2}}{n} = \frac{e_1^{\,2} + e_2^{\,2} + e_3^{\,2} + ...e_n^{\,2}}{n} \tag{8}$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{\sum e_t^{\,2}}{n}} = \sqrt{\frac{e_1^{\,2} + e_2^{\,2} + e_3^{\,2} + ...e_n^{\,2}}{n}} \tag{9}$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \left(\frac{1}{n}\sum\left|\frac{e_t}{y_t}\right|\right)(100) \tag{10}$$

$$= \left(\frac{\left|\frac{e_1}{y_1}\right| + \left|\frac{e_2}{y_2}\right| + \left|\frac{e_3}{y_3}\right| + ...\left|\frac{e_n}{y_n}\right|}{n}\right) \times 100$$

Where the forecast error is e_t , and it is calculated by subtracting the forecast value F_t from the series' actual value; y_t . Here n is the number of effective observations used to match the model. Minimum values of these accuracy measures provide the best fitting models. The lowest measurement of these error values shows and compares which methods are best suited to the model.

Findings and Discussions

Figure 1 shows the actual data on the daily cases of COVID-19 in Malaysia from July 2021 to December 2021 with a clear trend present. The highest number of new COVID-19 cases was 24599 on day 57, which ~~is~~ occurred on 26th August 2021. In contrast, the lowest number of reported COVID-19 cases from the data collected was 2589 cases on day 173, which ~~is~~ occurred on 20th December 2021. COVID -19 cases have shown a trend where there was an increase in the number of cases in August 2021. However, the number of COVID-19 cases started to drop from day 65. After day 128, the graph showed that the COVID-19 cases dropped to under 5000 cases and continued to decline.

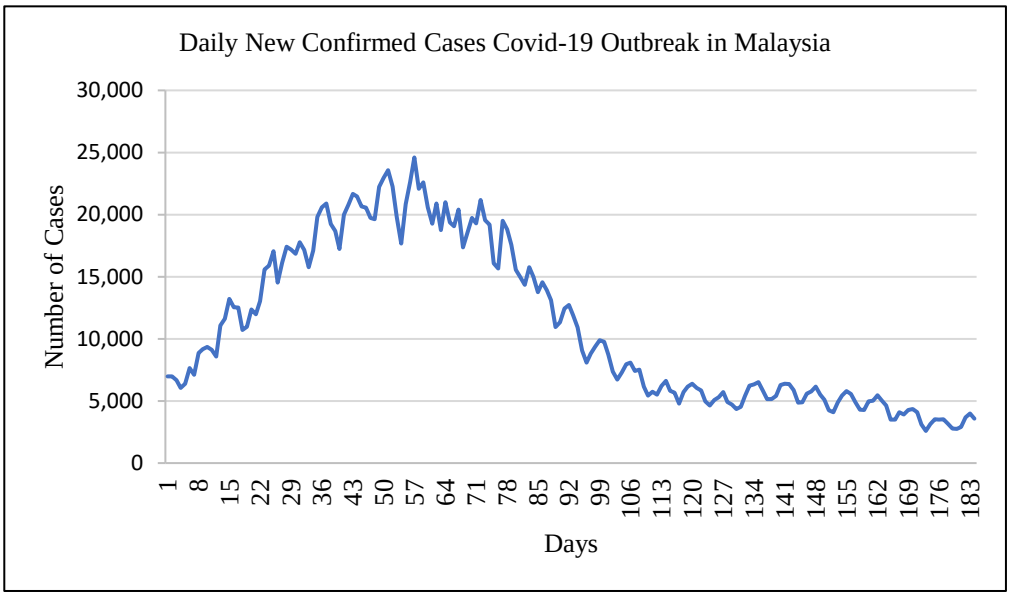

Figure 1: Daily New Confirmed Cases COVID-19 in Malaysia from July 2021 to December 2021

This dataset has a total of 184 data (N:184) from 1st July 2021 until 31th December 2021. The accuracy of the forecasted data was measured by comparing the generated models with the actual data. The ARIMA and fuzzy time series forecasting models were compared based on the lowest error measurements. Three types of error measurements were used in this study, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

In Table 1, the fuzzy time series model the lowest MSE, RMSE and MAPE measurement errors indicating higher accuracy in forecasting. The MSE of fuzzy time series and ARIMA (2,1,8) are 611279.3557 and 5167665.563 respectively. The RMSE of fuzzy time series is 781.8436 which is lower than ARIMA (2,1,8), 2273.250. The last measurements error evaluated is MAPE, fuzzy time series also has the lowest error measurement with the value of 4.3169 compared to ARIMA (2,1,8) with a value 53.629021. Therefore, the fuzzy time series is the best model chosen for predicting the number of new COVID-19 cases in Malaysia since it has lowest measurement errors.

Table 1
Measurement Errors for the Model Developed

Model	ARIMA (2,1,8)	Fuzzy Time Series
MSE	5167665.563	611279.3557
RMSE	2273.250	781.8436
MAPE	53.629021	4.3169

Figure 2 below shows the actual data of new cases COVID-19 and estimated COVID-19 using the Fuzzy time series method from July 2021 to December 2021.The graph shows that fuzzy time series method fit the actual data well.

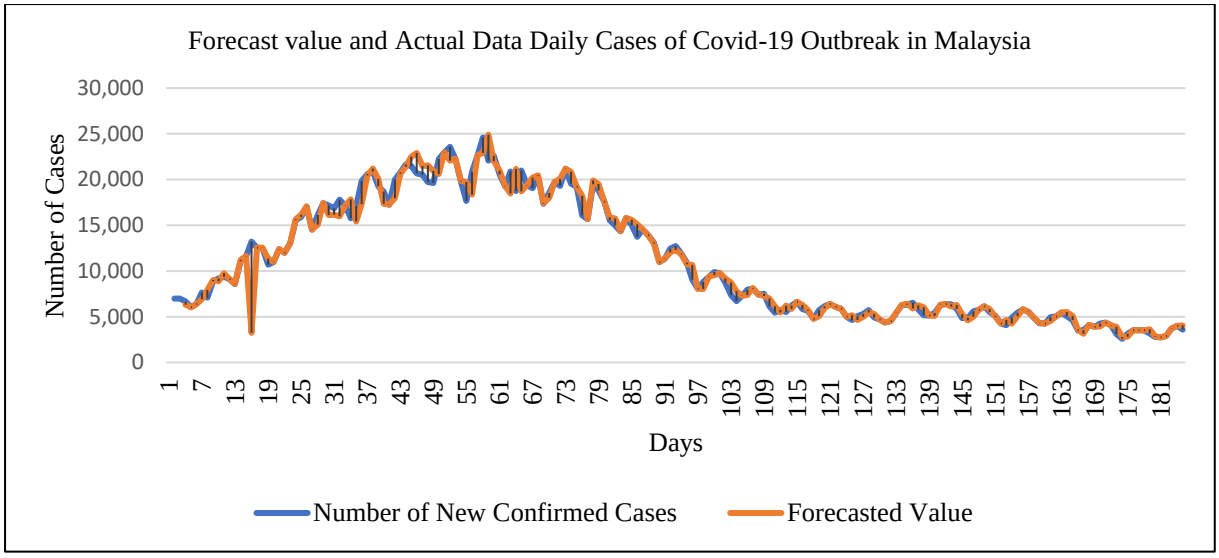

Figure 2: Forecast Value against Actual Daily New Confirmed Cases COVID-19 in Malaysia from July 2021 to December 2021

Conclusion and Recommendations

This study aimed to determine the optimal model for predicting COVID-19 case values in Malaysia. The evaluation was based on the minimum measurement errors of the model. Two methods, ARIMA and fuzzy time series, were used to analyze the actual COVID-19 cases data in Malaysia. The findings suggest that the fuzzy time series model was the best model for predicting the number of new COVID-19 cases in Malaysia, as it had the lowest measurement errors. This study's results can help the government in predicting future values of COVID-19 cases and assist in making informed decisions to combat the pandemic's spread.

Analysis of this study showed that the fuzzy time series is the best method since this model generated the smallest measurement error values compared to ARIMA (2,1,8). The graph of actual data of new COVID-19 cases and estimated COVID-19 cases from July 2021 to December 2021, showed that fuzzy time series method fits the actual data well. These results are consistent with previous studies by C. C. Wang (2011) who stated that fuzzy time series techniques behave more predictably than ARIMA time series techniques. In addition, Hadjira et al. (2021) revealed that fuzzy time series model is the best model in predicting COVID-19 outbreaks compared to ARIMA models using statistical criteria. For future studies, it is recommended to use more time series forecasting models and present the predictive values.

References

Alam, N. M. F. H. N. B., Ramli, N., & Nassir, A. A. (2022). Predicting Malaysian crude palm oil prices using Intuitionistic Fuzzy Time Series Forecasting Model. *ESTEEM Academic Journal*, *18*(March), 61–70.

Bernama. (2020). "First case of Malaysian positive for coronavirus", *Bernama*. Available: *https://www.bernama.com/en/general/news_covid-19.php?id=1811373*

Babulal V., and Othman N.Z. (2020, April 10). "Sri Petaling Tabligh gathering remains Msia's largest COVID-19 cluster", *New Straits Times*. Available: *https://www.nst.com.my/news/nation/2020/04/583127/sri-petaling-tabligh-gathering-remains-msias-largest-covid-19-cluster*

Cai, Q. Sen, Zhang, D., Wu, B., & Leung, S. C. H. (2013). A novel stock forecasting model based on fuzzy time series and genetic algorithm. *Procedia Computer Science*, *18*, 1155–1162.

https://doi.org/10.1016/j.procs.2013.05.281

Hadjira, A., Salhi, H., & El Hafa, F. (2021). A comparative study between ARIMA model, holt-winters – no seasonal and fuzzy time series for new cases of COVID-19 in Algeria. *American Journal of Public Health Research*, *9*(6), 248–256. https://doi.org/10.12691/ajphr-9-6-4

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Katoch, R., & Sidhu, A. (2021). An Application of ARIMA Model to forecast the dynamics of COVID-19 epidemic in India. *Global Business Review*. https://doi.org/10.1177/0972150920988653

Lee, D. H., Kim, Y. S., Koh, Y. Y., Song, K. Y., & Chang, I. H. (2021). Forecasting covid-19 confirmed cases using empirical data analysis in korea. *Healthcare (Switzerland)*, *9*(3). https://doi.org/10.3390/healthcare9030254

Mahmud, N., Syuhada Muhammat Pazil, N., Jamaluddin, H., & Aqilah Ali, N. (2021). Prediction of dengue outbreak: a comparison between ARIMA and Holt-Winters methods. *ESTEEM Academic Journal*, *17*(August), 101–111. https://ir.uitm.edu.my/id/eprint/6048/

Mishra, P., Fatih, C., Rawat, D., Sahu, S., Pandey, S. A., Ray, M., Dubey, A., & Sanusi, O. M. (2020). Trajectory of COVID-19 data in India: Investigation and project using artificial neural network, fuzzy time series and ARIMA models. *Annual Research & Review in Biology*, 46–54. https://doi.org/10.9734/arrb/2020/v35i930270

Moneim, H. A. A.-. (2020). Application of weighted fuzzy time series model to forecast epidemic njuries. *Current Journal of Applied Science and Technology*, *39*(24), 79–90. https://doi.org/10.9734/cjast/2020/v39i2430875

Pazil, N. S. M., Mahmud, N., Jamaluddin, S. H., & Ali, N. A. (2021). Forecasting dengue outbreak data using ARIMA model. *International Journal of Academic Research in Business and Social Sciences*, *11*(6). https://doi.org/10.6007/ijarbss/v11-i6/10106

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

Verma, P., Khetan, M., Dwivedi, S., & Dixit, S. (2020). Forecasting the COVID-19 outbreak: an application of ARIMA and fuzzy time series models. *Research Square*. https://doi.org/https://doi.org/10.21203/rs.3.rs-36585/v1

Wang, C. C. (2011). A comparison study between fuzzy time series model and ARIMA model for forecasting Taiwan export. *Expert Systems with Applications*, *38*(8), 9296–9304. https://doi.org/10.1016/j.eswa.2011.01.015

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., & Peng, Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA - Journal of the American Medical Association*, *323*(11), 1061–1069. https://doi.org/10.1001/jama.2020.1585

Wu, Y. C., Chen, C. S., & Chan, Y. J. (2020). The outbreak of COVID-19: An overview. In *Journal of the Chinese Medical Association* (Vol. 83, Issue 3, pp. 217–220). https://doi.org/10.1097/JCMA.0000000000000270.

WHO. (2020). Coronavirus disease (COVID-19): “Risks and safety for older people”. Available:

https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-risks-and-safety-for-older-people