Medical Science

pISSN 2321–7359; eISSN 2321–7367

Simple prediction accuracy for covid-19 daily incident data in Malaysia using machine learning- Auto Regressive Integrated Moving Average (ARIMA), and Linear Regression (LR) Model

Mohamad Arif Awang Nawi*, **Muhamamd Amirul Mat Lazin**

To Cite:
Nawi MAA, Lazin MAM. Simple prediction accuracy for covid-19 daily incident data in Malaysia using machine learning- Auto Regressive Integrated Moving Average (ARIMA), and Linear Regression (LR) Model. Medical Science 2022; 26:ms292e2305.
doi: https://doi.org/10.54905/disssi/v26i125/ms292e2305

Authors' Affiliation:
School of Dental Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia.

***Corresponding Author**
School of Dental Sciences, Health Campus, Universiti Sains Malaysia,
16150 Kubang Kerian, Kelantan,
Malaysia
Email: mohamadarif@usm.my

Peer-Review History
Received: 18 May 2022
Reviewed & Revised: 20/May/2022 to 15/July/2022
Accepted: 18 July 2022
Published: 20 July 2022

Peer-review Method
External peer-review was done through double-blind method.

URL: https://www.discoveryjournals.org/medicalscience

DISCOVERY
SCIENTIFIC SOCIETY

ABSTRACT

Introduction: The COVID-19 pandemic has spread so widely across borders and worldwide. All countries, including Malaysia, need to take immediate and aggressive action to curb and combat this virus by accurately predicting the number of new COVID-19 cases a few weeks in advance. This paper aims to present robust traditional methods that can be used to make accurate predictions for the evolution of an epidemic over the next two weeks. *Methodology:* There are two traditional methods for forecasting, such as ARIMA and LR. We use COVID-19 daily data in Malaysia from 1 January 2021 to 14 January 2022, and use Microsoft Excel, Minitab, and SPSS software version 26 to perform the forecasting procedure. *Results:* The ARIMA model showed a slight difference between the forecast value and the actual value of the new COVID-19 case compared to the forecast value for the LR model. This indicates that the efficient and usable prediction model is the ARIMA model with high adj-R^2 values and low MAD, MSE, RMSE, and MAPE values. *Conclusions:* Accurately predicting trends is an important aspect of preventing the spread of COVID-19 outbreaks, especially in countries with large populations. Furthermore, accurate forecasts can provide feedback on whether the policies implemented effectively relieve the pressures of the national healthcare system and enable the government to evaluate different strategies for risk reduction and to regulate different policies based on projections of different areas of concern.

Keywords: Prediction, COVID-19, ARIMA, LR, Malaysia

1. INTRODUCTION

The pandemic of COVID-19 has impacted people's social and economic lives worldwide. It has presented government agencies with a major public health challenge. When the COVID-19 pandemic occurs, decision-makers face the difficulty of uneven development and the model's limitations in predicting its evolution—the rising number of confirmed cases raised the demand for ICU beds in the worst situations. Monitoring and predicting the new cases are critical because of infection spread that shows significant variation at the local level. If local decision-makers can accurately predict the number of new COVID-19 cases a few weeks in advance, they can take all the steps needed to make sure there are enough hospital beds to treat COVID-19 patients before they occur rather than responding to it (Rostami-Tabar and Rendon-Sanchez, 2021).

There are many benefits to accurately predicting the number of new COVID-19 cases. First, it gives public health agencies a better way to plan for new cases in the future. Second, a prediction of a rapid rise in the number of new cases would let government officials take the necessary steps to stop the virus from spreading quickly. A forecast of fewer new cases, on the other hand, would allow officials to avoid imposing unnecessarily strict restrictions. Accurate predictions would help officials make better decisions in general. As a result, it's critical to choose a method that can accurately predict incidence (Kamalov and Thabtah, 2021). Forecasting methods have been used to look for patterns and trends during major epidemics like the Avian Flu and Ebola. Epidemic trajectory and the benefits of various intervention techniques to slow disease spread were predicted using epidemic models developed for these disease conditions. Model selection is based on performance indicators like Mean Square Error (MSE), Mean Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE). Therefore, the lowest value of MSE, MAD, and MAPE is the "best" model (Brockwell and Devis, 2016).

ARIMA time series approach has been utilized in research to predict incidence cases (Benvenuto et al., 2020). ARIMA and exponential smoothing are two of the most common methods (ETS). Additionally, automatic elements have been included in statistical tools such as R Studio and IBM SPSS to fit the best prediction model. ARIMA consists of three components such as autoregressive (p), differencing (d), and moving average (q) (Chintalapudi et al., 2020). These three components provide the platform for better predictive modelling. In addition, ARIMA has the advantage of including regressive predictors to improve forecasting accuracy. However, the model is prone to over fit and less appropriate if careful modelling considerations are not taken as limited historical data or time points (Hyndman and Athanasopoulos, 2018).

This paper aims to present a robust traditional method that can be used to reliably make predictions for the evolution of the epidemic over the next two weeks. The number of new COVID-19 cases in Malaysia is predicted using ARIMA and Linear Regression. Additionally, we compare the models based on the performance metrics used for model selection, such as MAD, MSE, RMSE, and MAPE.

2. METHODOLOGY

"How accurate are traditional, quantitative forecasting approaches at predicting the occurrence of COVID-19 in Malaysia?" is our research question. Data from Malaysia is chosen because, first, there's accessibility. Malaysia was one of the first locations to receive complete data due to the incidence occurring before other country sections. Second, the data revealed a complete curve: an upward trend, a peak, and a downward trend. The study assessed a variety of forecasting methods, such as ARIMA and LR. We tested COVID-19 daily incidence data in Malaysia from January 1, 2021, to January 14, 2022, from a daily press release issued by the Ministry of Health Malaysiaon COVID -19. Microsoft Excel, Minitab, and SPSS software version 26 were used to execute the forecasting procedures. The differences between the actual and the forecast values for each forecasting method were determined, and the lowest MAD was calculated.

Data Pre-processing

After the data pre-processing, ARIMA and linear regression are implemented. Finally, the performances of the two models are compared and forecasted for the coming 14 days.

Linear Regression (LR)

LR is the most basic supervised machine learning algorithm and is widely used to form forecasting models. This analysis is best suited for case representation, and then linear trend prediction of COVID-19 outbreaks will be available for forecasting purposes (Nawi, 2020). Therefore, estimating the number of new COVID-19 cases based on death cases is important. The linear trend line is best suited to the recent case according to the pattern search. The basic prediction equation is $y = mx + b$, where m is the slopand b is the y-intercept, which expresses the linear relationship between the independent variable (x) and the dependent variable (y). The number of new COVID-19 positive cases and the number of death cases in Malaysia were the variables in the study.

ARIMA

An ARIMA model specification consists of $''p''$ refers to the autoregressive order (AR) component, $''d''$ refers to the differencing degree of the original time series, and $''q''$ refers to the moving average (MA) component (Cao et al., 2013). Model-identification, parameter estimate, and model validation were the criteria that needed to be taken to select an appropriate ARIMA model.The unknown coefficients for the components of AR and MA were determined using multivariate regression analysis in the parameter estimation step (Beaumont et al., 1984). The processes outlined above are conventional in ARIMA modelling methodology, and all phases use the SPSS statistical software. The validation stage determines which model is the most appropriate for the upcoming data by comparing them. The ability of the ARIMA model with covariates to match predictive data with actual COVID-19 case datawill serve as the final test, and it will be accompanied by a confidence interval of 95%.

Evaluation metrics

Small and random discrepancies between actual and predicteddata (called "residuals") indicate that a model has a good fit to real-world data. This is called "excellent fit." Models with the lowest MAD, MSE, RMSE, and MAPE produced the most accurate forecasts (figure 1).

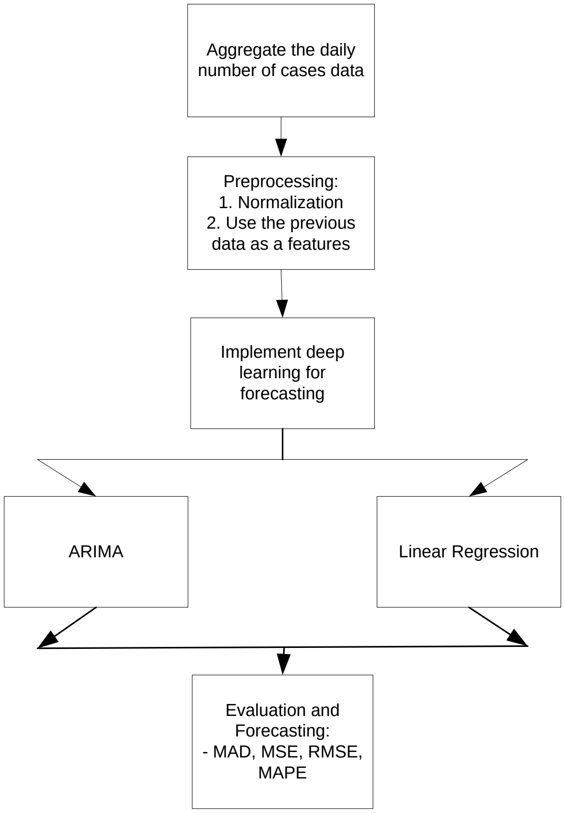

Aggregate the daily number of cases data
Preprocessing:
1. Normalization
2. Use the previous data as a features
Implement deep learning for forecasting
ARIMA
Linear Regression
Evaluation and Forecasting:
- MAD, MSE, RMSE, MAPE

Figure 1 Flowchart of the Study

3. RESULTS

ARIMA

During the model estimation training period, the ACF and PACF that were made by the 5-point MA smoother for the COVID-19 cases did not gradually decrease across the different lag points. This shows that the series is not stationary. Therefore, as shown in Figure 2, differencing was justified to transform the series to stationarity.

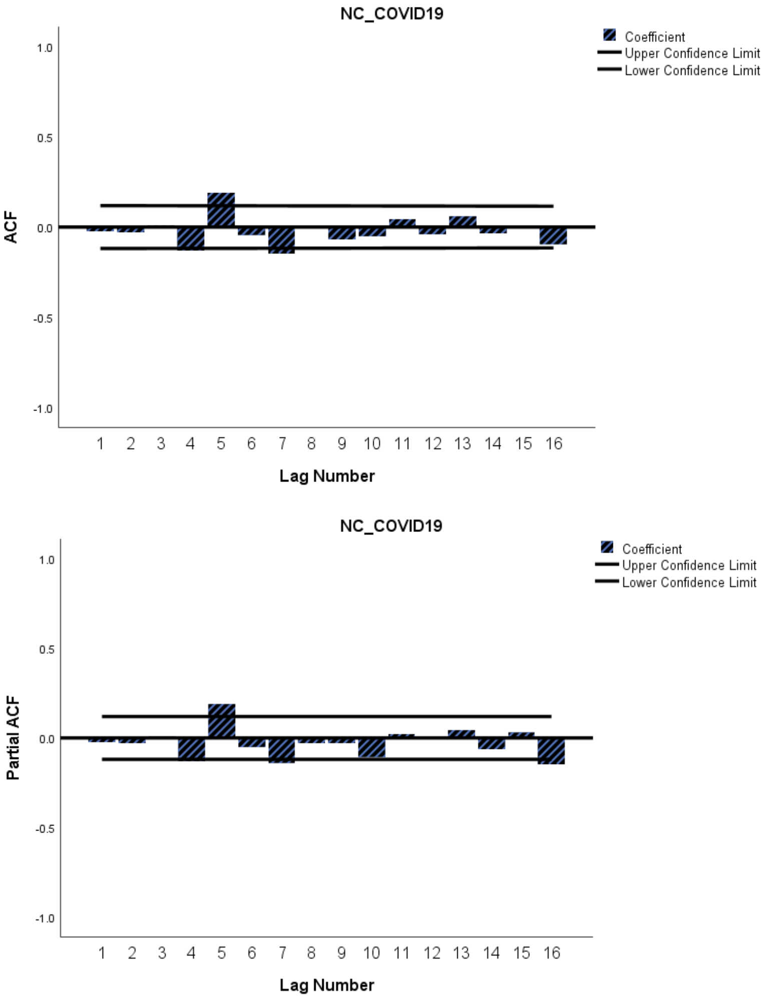


Figure 2 Daily new cases of COVID-19 according to first order difference ACF and PACF plots in Malaysia

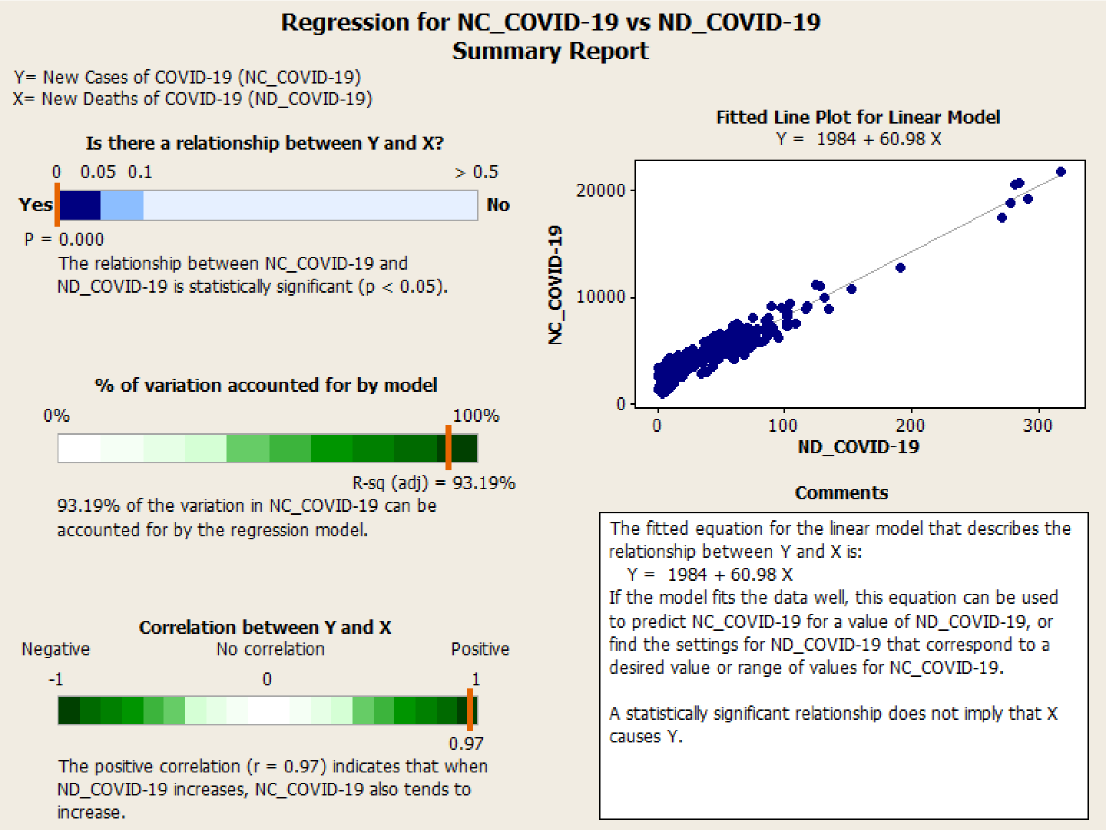
The ARIMA model (p, d, q) will be formed with orders (1, 0, 1) where 1 is the order of AR model, 0 is differencing data, and 1 is MA.

Linear regression

Mathematical equations that describe the linear regression model between new COVID-19 positive cases and death cases of COVID-19 are shown below:

$Y = 1984 + 60.98\ X$

This equation can be used to predict a new daily COVID-19 case against a COVID-19 death case if the model fits the data. The existence of a statistically significant relationship does not necessarily imply that the variable x is the driving force behind the variable y (figure 3).



Regression for NC_COVID-19 vs ND_COVID-19
Summary Report

Y= New Cases of COVID-19 (NC_COVID-19)
X= New Deaths of COVID-19 (ND_COVID-19)

Is there a relationship between Y and X?

0 0.05 0.1 > 0.5

Yes No

P = 0.000
The relationship between NC_COVID-19 and ND_COVID-19 is statistically significant (p < 0.05).

% of variation accounted for by model

0% 100%

R-sq (adj) = 93.19%
93.19% of the variation in NC_COVID-19 can be accounted for by the regression model.

Correlation between Y and X

Negative No correlation Positive

-1 0 1

0.97
The positive correlation (r = 0.97) indicates that when ND_COVID-19 increases, NC_COVID-19 also tends to increase.

Fitted Line Plot for Linear Model
Y = 1984 + 60.98 X

Comments

The fitted equation for the linear model that describes the relationship between Y and X is:
Y = 1984 + 60.98 X
If the model fits the data well, this equation can be used to predict NC_COVID-19 for a value of ND_COVID-19, or find the settings for ND_COVID-19 that correspond to a desired value or range of values for NC_COVID-19.

A statistically significant relationship does not imply that X causes Y.

Figure 3 Summary Report for Linear Regression (LR) for New Cases of COVID-19 vs Death cases of COVID-19 in Malaysia

Comparison between ARIMA and LR model

Figure 4 illustrates the number of new COVID-19 cases for three different readings. The forecast period runs from January, 1 2022 to January, 14 2022 (for 14 days). The blue line indicates the actual value of the COVID-19 positive case, the orange line indicates the forecast value for the ARIMA model, and the grey line indicates the forecast value for the LR model with a 95%confidence level. It is clear from the graph that the number of new COVID-19 cases recorded the highest number on 6 January 2022 (3543 cases) and 13 January 2022 (3684).

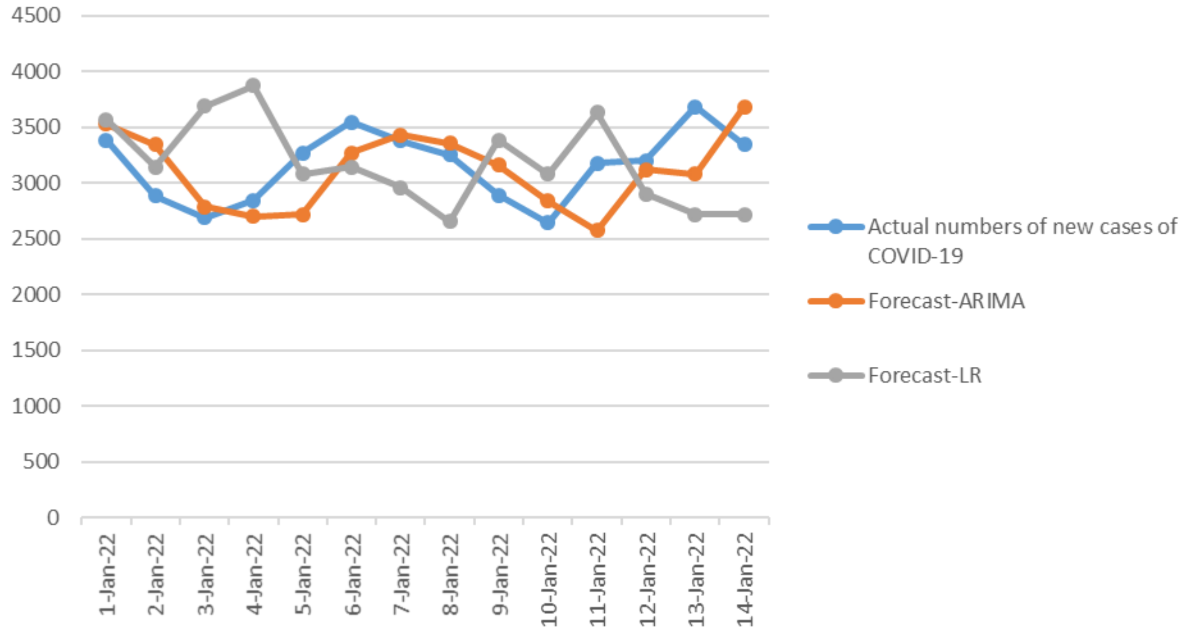


Figure 4 Forecasting values of New Cases of COVID-19 for next 14 days.

If we look at the results of the ARIMA model, there is a slight difference of the forecast value and the actual value for the new COVID-19 case compared to the forecast value for the LR model (Table 1). This indicates that an efficient and usable forecasting model is the ARIMA model.

Table 1 The number of New Cases of COVID-19 for Actual and Forecast using ARIMA and LR

Date	Actualnumbers of new cases of COVID-19	Forecast-ARIMA	Forecast-LR
01-Jan-2022	3386	3528	3569
02-Jan-2022	2882	3342	3143
03-Jan-2022	2690	2788	3691
04-Jan-2022	2842	2703	3874
05-Jan-2022	3270	2715	3082
06-Jan-2022	3543	3268	3143
07-Jan-2022	3381	3433	2960
08-Jan-2022	3251	3356	2655
09-Jan-2022	2888	3161	3386
10-Jan-2022	2641	2841	3082
11-Jan-2022	3175	2575	3630
12-Jan-2022	3198	3122	2899
13-Jan-2022	3684	3081	2716
14-Jan-2022	3346	3684	2716

According to the adj-R^2, the ARIMA model performed best for this scenario which is shown below in Table 2. The ARIMA model was selected with the number of death COVID-19 cases as independent covariates and the most suitable model with the lowest MAD, MSE, RMSE and MAPE compared to the LR model (Table 2).

Table 2 Model Fit Statistics between ARIMA and LR

Model	ARIMA	LR
Adj-R^2	0.945	0.932
MAD	535.15	686.50
MSE	20.27	24.39
RMSE	410.87	595.05
MAPE	12.40	22.34

4. DISCUSSION

Tandon et al., (2020) analysed time series data of validated COVID-19 cases in India using the ARIMA model.First, the ARIMA parameters were determined using an ACF and a PAF graph. According to Dehesh et al., (2020), an ARIMA analysis of multiple countries shows that more stringent intervention results in a stationary incidence trend. To get the best prediction based on the type of outcome chosen, manual ARIMA requires a more selective model and approach. It was recently reported that European countries that use total cumulative cases forecast future cases differently than countries that predict new daily cases (Ceylan, 2020). As a result, any forecasting is data-driven, and interpretation requires caution.

Short-term forecasting with a seven-day forecasting period aids policymakers in making informed public health decisions. For example, MCO was announced just a few days before the two-week period ended. For starters, it marks the halfway point of the incubation period, allowing asymptomatic cases to be screened before they manifest symptoms. Second, the seven-day doubling time has been suggested in the literature as a signal for rapid public health intervention to reduce transmission (Salim et al., 2020).

5. CONCLUSION

This study uses machine learning to predict the positive number of COVID-19 for the next two weeks in Malaysia using ARIMA and LR methods. The results obtained from this study illustrate that the ARIMA model is the most efficient based on the model fix value. Thus, forecasting the COVID-19 situation-related model can help the government know the future situation. However, the results of the essential attributes and predictions depend on the reliability of the data set. In addition, the effectiveness of ARIMA in modelling the case spread of COVID-19 may differ for different data samples. Therefore, understanding the spread and predicting trends accurately is important criteria of preventing the spread of an epidemic, especially in countries with large populations. The government needs to make accurate predictions on the spread of COVID-19 so that each country can prepare in the face of the epidemic. Furthermore, accurate predictions can provide feedback on whether the policies implemented effectively relieve the stress of the national health care system.

Malaysia currently has a shortage of hospital beds and estimating the number of beds is much needed to effectively predict an increase or decrease in the number of new COVID-19 positive cases. Furthermore, this research will aid the government and the Malaysia Strategic Work plan for Emerging Disease (MySED Work plan) in preparing for a future surge in confirmed cases. However, the various factors influencing a confirmed case make it difficult to predict how it will change accurately. Furthermore, mass inflection has a significant impact. Overall, research focusing on various techniques such as machine learning reinforcement, deep learning modelling and predictive algorithms is needed to curb this epidemic from spreading more quickly and in turn reduce the recorded casualties.

Acknowledgments

The authors would like to express their gratitude to Universiti Sains Malaysia (USM) for providing the research funding (Short Term Grant No.304/ PPSG/6315410, School of Dental Sciences, Health Campus, Kelantan, Malaysia).

Author Contributions

Mohamad Arif Awang Nawi: Contributed in the drafting of the manuscript, conception and design of the study, analysis, interpretation of data and final approval.

Muhamamd Amirul Mat Lazin: Contributed in acquisition of data, analysis, interpretation of data, and final approval.

Funding

This study was funded by Universiti Sains Malaysia (Short Term Grant No.304/PPSG/6315410).

Conflicts of interest

The authors declare that there are no conflicts of interests.

Data and materials availability

All data associated with this study are present in the paper.

REFERENCES AND NOTES

1. Beaumont C, Makridakis S, Wheelwright SC, McGee VE. Forecasting: Methods and applications. J Oper Res Soc 1984; 35: 79-81
2. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief 2020; 26:105340. doi: 10.1016/j.dib.2020.105340

3. Brockwell PJ, Davis RA. Introduction to time series and forecasting. Springer 2016
4. Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, Lu Z. A hybrid seasonal prediction model for tuberculosis incidence in China. BMC Med Informat Decision Making 2013; 13:56. doi: 10.1186/1472-6947-13-56.
5. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020; 729:138817. doi:10.1016/j.scitotenv.2020.138817
6. Chintalapudi N, Battineni G, Amenta F. COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. J Microbiol Immunol Infect 2020; 53(3):396-403. doi: 10.1016/j.jmii.2020.04.004
7. Dehesh T, Mardani-Fard HA, Dehesh P. Forecasting of covid-19 confirmed cases in different countries with ARIMA models. Med Rxiv 2020.doi:org/10.1101/2020.03.13.20035345
8. Hyndman, RJ, Athanasopoulos G. Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia 2018. Accessed on 2022 March 8
9. Kamalov F, Thabtah F. Forecasting Covid-19: SARMA-ARCH approach. Health Technol 2021; 11: 1139–1148. doi: 10.1007/s12553-021-00587-x
10. Nawi MAA. Malaysia's Pattern of COVID-19 Cases from 15 February until 14 March 2020 and Prediction of Deaths Cases in Phase 3 of Movement Control Order (MCO). Int Med J 2020; 27(3): 254 – 258
11. Rostami-Tabar B, Rendon-Sanchez JF. Forecasting COVID-19 daily cases using phone call data. Appl Soft Comput 2021; 100:106932. doi: 10.1016/j.asoc.2020.106932
12. Salim N, Chan WH, Mansor S, Bazin NE, Amaran S, Faudzi AA, Zainal A, Huspi SH, Khoo EJ, Shithil SM. COVID-19 epidemic in Malaysia: Impact of lock-down on infection dynamics. medRxiv 2020. doi: 10.1101/2020.04.08.20057463
13. Tandon H, Ranjan P, Chakraborty T, Suhag V. Coronavirus (covid-19): Arima based time-series analysis to forecast near future. medRxiv 2020. arXiv:2004.078592020