JIDC | THE JOURNAL OF INFECTION IN DEVELOPING COUNTRIES

Coronavirus Pandemic

Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models

Sarbhan Singh[1], Bala Murali Sundram[1], Kamesh Rajendran[1], Kian Boon Law[2], Tahir Aris[1], Hishamshah Ibrahim[3], Sarat Chandra Dass[4], Balvinder Singh Gill[1]

[1] *Institute for Medical Research (IMR), Ministry of Health, Kuala Lumpur, Malaysia*
[2] *Institute for Clinical Research (ICR), Ministry of Health, Shah Alam, Malaysia*
[3] *Ministry of Health, Putrajaya, Malaysia*
[4] *Heriot-Watt University, Putrajaya, Malaysia*

Abstract

Introduction: The novel coronavirus infection has become a global threat affecting almost every country in the world. As a result, it has become important to understand the disease trends in order to mitigate its effects. The aim of this study is firstly to develop a prediction model for daily confirmed COVID-19 cases based on several covariates, and secondly, to select the best prediction model based on a subset of these covariates.
Methodology: This study was conducted using daily confirmed cases of COVID-19 collected from the official Ministry of Health, Malaysia (MOH) and John Hopkins University websites. An Autoregressive Integrated Moving Average (ARIMA) model was fitted to the training data of observed cases from 22 January to 31 March 2020, and subsequently validated using data on cases from 1 April to 17 April 2020. The ARIMA model satisfactorily forecasted the daily confirmed COVID-19 cases from 18 April 2020 to 1 May 2020 (the testing phase).
Results: 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 forecasted values showed a downward trend of COVID-19 cases until 1 May 2020. Observed cases during the forecast period were accurately predicted and were placed within the prediction intervals generated by the fitted model.
Conclusions: This study finds that ARIMA models with optimally selected covariates are useful tools for monitoring and predicting trends of COVID-19 cases in Malaysia.

Key words: COVID-19; ARIMA; forecast; pandemic.

J Infect Dev Ctries 2020; 14(9):971-976. doi:10.3855/jidc.13116

(Received 23 May 2020 – Accepted 13 July 2020)

Introduction

On 11 March 2020, the World Health Organization (WHO) declared COVID-19 a pandemic. Since it was first discovered in Wuhan, China, on 31 December 2019, this pathogen has rapidly spread and have infected more than 4 million people globally, with over 300,000 deaths as of March 2020 [1]. This virus is highly transmissible with a basic reproductive number (R_0) estimated between 2.5 to 3.5 [1]. This disease has high case fatality rates especially among the elderly population with underlying co-morbidities that ranges between 2% to 18% [2]. With no specific treatment, cure or vaccine available for COVID-19, the challenges facing healthcare systems worldwide to manage this pandemic is enormous.

Since COVID-19 was first detected on 25 January 2020 in Malaysia, over 6,000 people have been infected by the virus, and a total 113 deaths has been reported as of 18 May 2020 [3]. In order to curb this pandemic, Malaysia has instituted several non-pharmaceutical intervention (NPI) strategies including the Movement Control Order (MCO) which was first implemented on 18 March 2020. The MCO along with the other measures such as social distancing, isolation and quarantine aimed to break the chain of transmission of COVID-19 in Malaysia. These measures were implemented with the aim to flatten the epidemic curve and prevent an exponential rise in new COVID-19 infections, that would allow for the effective management and control of the pandemic.

Accurate forecasting of COVID-19 case trends is essential for the preparedness of health systems in terms of outbreak management and resource planning. Mathematical and statistical modelling of infectious disease are effective tools that would enable health systems to anticipate future disease trends. Time series

models such as the Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) have been widely used to statistically model and forecast infectious disease trends [4]. ARIMA models are preferred in this context as they are suitable for investigations into short-term effects of acute infectious diseases and are a flexible class of models that are appropriate to fit several trajectories, and have been well documented in the literature [4,5]. Furthermore, in the current situation, ARIMA models are preferred over their SARIMA counterparts since the data points span a period of less than one year, negating the effect of any seasonal variation. ARIMA models have been used in several studies in Italy, Brazil, China, Iran, Thailand and South Korea to forecast the COVID-19 outbreak trends. However, to date, there are no studies conducted using ARIMA models to forecast the COVID-19 outbreak in Malaysia [6–8]. In this study, ARIMA models were developed using daily COVID-19 confirmed and active cases in Malaysia to identify the best fitting model to forecast COVID-19 cases from 18 April 2020 to 1 May 2020. Forecasting future COVID-19 cases using ARIMA models are suitable especially when model parameters that determine the disease dynamics are unavailable or undetermined due to the disease novelty. In addition, ARIMA models are a flexible, empirical method which is able to produce reliable forecast in situations with limited data. This paper demonstrates that, ARIMA models are able to provide reasonable forecasts even with the above mentioned limitations.

Methodology

Data Source

Data for this study were obtained from 22 January 2020 to 1 May 2020. The Malaysian daily COVID-19 confirmed and active cases were sourced from MOH Malaysia's official website, http://www.moh.gov.my/, and from MOH daily press releases. Daily COVID-19 confirmed cases for neighboring countries were also obtained from the Johns Hopkins University's official website (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html). Statistical Package for the Social Sciences (SPSS) version 24.0 was used to develop the time-series database and prediction models.

ARIMA model development

The ARIMA model with covariates was used to predict the trend of daily confirmed COVID-19 cases from 18 April to 1 May 2020. Case data from 22 January 2020 to 17 April 2020 were used for model training and validation purposes. The training and validation period used 87 data points which was more than the minimum requirement of 50 data points recommended to provide reliable forecasts [9]. ARIMA models were then fitted to the observed time series data. The training-testing framework was utilized to validate the ARIMA model and determine the accuracy of the model fit and forecasts. We found that the accuracy of prediction was reliable in this framework and hence we extended the model fit using all 87 data points to produce our final forecast for unobserved future cases. The general specification of an ARIMA model is as follows: (p, d, q) where p refers to the order of autoregressive (AR) component, d refers the degree of differencing of the original time series, and q refers to the order of the moving average (MA) component [10-11].

The determination of a suitable ARIMA model involved three stages – model identification, parameters estimation and model validation. Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the original time series data were obtained to check for stationarity, and to provide a specification of the p, d, and q parameter terms. In the parameter estimation step, the unknown coefficients corresponding to the AR and MA components of the ARIMA model were estimated by implementing a multivariate regression analysis [12]. The above procedures are standard in the methodology of ARIMA modelling and the SPSS statistical software was used to carry out all the steps.

The daily COVID-19 cases were found to show a high degree of variability (noise) and skewedness which was determined by estimating the variance and test of normality. To remove the noise component, the data was averaged on a moving window using a 5-point MA smoother. The 5-point MA smoother was found to have a lower variance over the 87 data points while still maintaining the general observed trend of the original time series. Hence, the 5-point MA smoother was chosen subsequently as the dependent variable for the forecast model (see Supplementary Figure 1). The covariates considered for developing the prediction model were selected from (1) daily active, (2) state and (3) district COVID-19 cases in Malaysia. In addition, (4) daily COVID-19 cases for Singapore was also considered as a covariate. Thus, there was a total of 4 independent covariates considered in this study.

The entire time series data was divided into two parts: a training set period from 22 January 2020 to 31 March 2020 to fit the models, and a validation period

from 1 to 17 April 2020 to validate the fitted models. The validation stage was carried out to determine which model fits future data sufficiently well. A model is considered to be a good fit to the observed data if the differences between the observed and predicted cases (i.e., the residuals) are small and random [13]. Models with a lower Bayesian Information Criterion (BIC) and Mean Absolute Percentage Error (MAPE) were considered the best fitting models to produce the most accurate forecasts. The Ljung-Box test was performed on the residuals to determine if the residual ACF at different lag times was significantly different from zero as further evidence of the model fit [12].

The same process as described above was used to predict the values of each of the four covariates for the forecast period from 18 April 2020 to 1 May 2020. The time series for these covariates from 22 January 2020 to 1 May 2020 was then used to develop the best fitting model to forecast the 5 - point MA COVID-19 cases from 18 April 2020 to 1 May 2020 in Malaysia. All covariates were found to have no evidence of multicollinearity by using the variance inflation factor (VIF) test [13]. The final test for the ARIMA model with covariates was its ability to fit the forecast data to the actual observed COVID-19 cases during the forecast period. The 95% confidence interval (CI) of the fit and forecast values were also generated and verified. In the event the lower CI values were less than 0, the negative values were converted to 0 to avoid negative COVID-19 case projections [13]. Differences between the daily observed and forecasted cases was estimated using a deviation index calculated by the following formula.

$$Index = \frac{5\ point\ moving\ average\ for\ observed\ case - 5\ point\ moving\ average\ for\ Forecast\ case}{5\ point\ moving\ average\ for\ Forecast\ case} \times 100$$

Results

In the training period for the model estimation, the generated ACF and PACF for the 5-point MA smoother

Figure 1. ACF and PACF for the ARIMA (0,1,0).

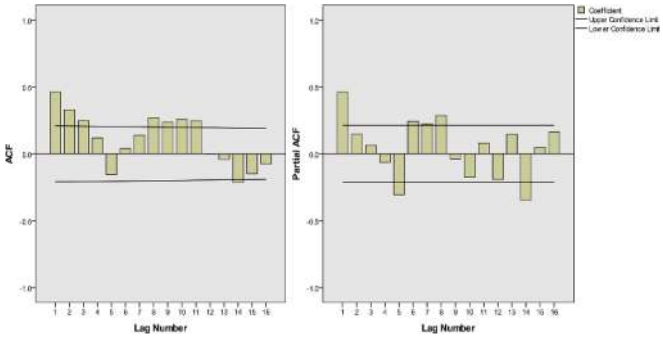
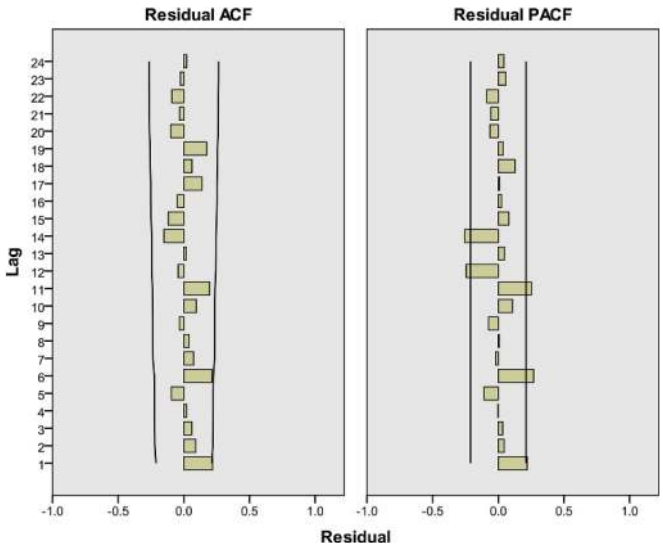


Figure 2. ACF and PACF residuals for the ARIMA (0,1,0).



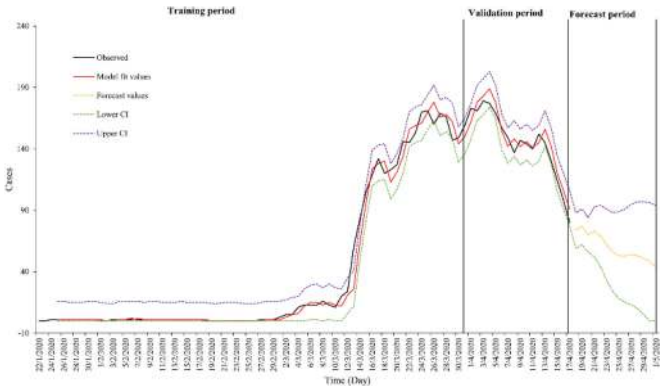
for the COVID-19 cases did not appear to be gradually tailing off across the different lag points therefore indicating a non-stationary series. Differencing was justified to transform the series to stationarity as shown in Figure 1. In addition, the Augmented Dicky Fuller test (ADF) for the difference data set at lag 1 subsequently showed that the data set was stationary (ADF test statistic was -3.75, p = 0.01) [11]. A total of 5 models were developed using various combinations

Table 1. MAPE and normalized BIC for ARIMA models.

Model	MAPE	BIC	Q18	p-value
ARIMA (0,1,0) with covariates – Active cases State and district cases*	16.014	4.172	23.115	0.186
ARIMA (1,1,5)	21.960	4.049	36.123	0.003
ARIMA (1,1,5) with covariates – Active cases	21.211	4.075	29.240	0.022
ARIMA (1,1,0) with covariates – Active cases Neighboring country**	16.319	4.158	50.111	0.000
ARIMA (1,1,0) with covariates – Active cases State and district cases* Neighboring country**	16.571	4.073	37.593	0.003

* Selangor state and Lembah Pantai district; ** Singapore.

Figure 3. Model validation and forecast of ARIMA (0,1,0) for period from 22 January 2020 – 1 May 2020, Malaysia.



of independent covariates as shown in Table 1. The ARIMA (0,1,0) model with active cases, state and district cases as independent covariates was selected as the best fit model with the lowest MAPE = 16.014 and BIC = 4.172. In addition, the model was validated using the Ljung-Box Q(18) test, which suggested that the ACF for the residuals at different lag times was not statistically different from zero (Ljung-Box Q(18) test = 23.115, p = 0.186). Analysis of the ACF and PACF residuals for the ARIMA (0,1,0) shows that the model fits the data reasonably well, with no residuals falling outside the 95% CI (Figure 2). Therefore, the ARIMA (0,1,0) model using active cases, state and district cases as independent covariates produced the best fitting model as shown in Figure 3. The estimated coefficients in the fitted ARIMA (0,1,0) are shown in Table 2.

As shown in Figure 3, the ARIMA (0,1,0) forecast showed a downward trend of COVID-19 cases during the forecast period from 18 April 2020 to 1 May 2020. In addition, the 5-point MA values for the COVID-19 cases were within the 95% CI prediction intervals as shown in Figure 4. All forecast values were within a 25% deviation range from the smoothen observed cases with 80% of forecast values within the 15% deviation index as shown in Figure 5. The model predicted the cumulative COVID-19 cases accurately during the forecast period (18 April 2020 to 1 March 2020) where there were a total of 837 cases predicted compared to 820 cases that was observed. This was a difference of only 17 cases higher than the observed (1.9%).

Figure 4. Observed and forecast COVID-19 daily cases for ARIMA (0,1,0) for period 18 April 2020 to 1 May 2020, Malaysia.

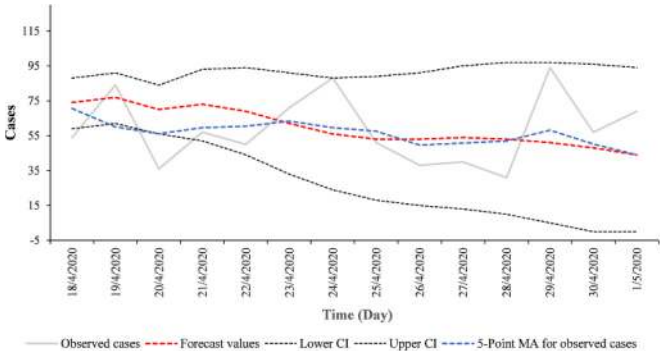


Figure 5. Deviation index of the ARIMA (0,1,0) forecast from 18 April 2020 to 1 May 2020, Malaysia.

Discussion

Since the WHO declared COVID-19 a pandemic on 11 March 2020, several countries including Malaysia experienced an exponential rise in COVID-19 cases [14]. This rapid increase of cases has stressed most healthcare systems worldwide and has further made outbreak response and resource planning a challenge. In response, health authorities have attempted to forecast the trend of this pandemic, however this have proven to be difficult as COVID-19 is a novel disease with limited data and knowledge on the disease trends and dynamics [15]. This is especially observed when using compartmental and time series models to predict disease trends, where compartmental models such as the SIR/SEIR models requires sufficient data to

Table 2. Model parameters for ARIMA (0,1,0).

Model covariates	Estimate	SE	p-value
Constant	2.079	1.140	0.072
State case	0.111	0.069	0.112
District case	-0.005	0.001	0.001
Active case	0.105	0.088	0.234

determine model parameters, and ARIMA models require long time series data to be accurate. However, ARIMA models may be preferred to compartmental models when there is insufficient data on the disease dynamics to generate the required parameters for model estimation [9]. Time series ARIMA models are considered more suitable with long time series data, however, with as little as 50 data points, in these models are able to provide reliable forecasts [9].

During the COVID-19 pandemic where there is limited data and information on the disease, we found that time series ARIMA models was able to forecast the COVID-19 case trends from 18 April 2020 to 1 May 2020 accurately with observed cases being well within the 95% CI of the forecast. Our forecast also showed an accurate reducing trend which corresponded to the observed cases from 18 April 2020 to 1 May 2020. As shown in Figure 5, this finding is strengthened by variations of less than 15% between the forecast and observed cases in 80% of the forecasted data points. This paper demonstrates that ARIMA models are a suitable tool to forecast case trends especially during situations where data is limited. Similarly, studies on COVID-19 conducted in countries such as South Korea, Iran and Italy was able to predict case trends using ARIMA models in similar conditions [7]. In addition as with our findings, a study in Italy also reported a high level of forecast accuracy of 95% in predicting COVID-19 trends using ARIMA models [8].

This ability for the model to generate an accurate forecast with limited data can be attributed to several reasons which includes the preparation of the dependant variable before model estimation and incorporating suitable covariates into the model. As this study was using daily data, it was observed to be highly variable and inherently noisy. Using smoothened data have shown to increase model accuracy especially when limited number of data points are available [16-18], hence, the dataset was first smoothened to remove skewedness and noise.

In addition, improving trend signals of noisy data can be achieved by the use of independent covariates that have well-defined trends. In this study we used active case data that had less variability (noise) to provide clearer trend signals in the model development, which subsequently increases the model fit. The use of subnational data with larger outbreaks and high case numbers as independent covariates was able to provide clear trend signals and improve the model output for Malaysia which would otherwise have been obscured by national level data. These findings has strengthened the use of independent covariates which increases the model accuracy and reliability in situations where data is limited, as it allows for the detections of signals and changes in case trends accounted by suitable covariates as were also reported in previous studies [16].

The strengths of this study include, firstly, this paper is the first to report the use of ARIMA models to forecast COVID-19 cases and trends in Malaysia. Secondly, this was the first attempt to use smoothen case data to reduce noise and improve accuracy as compared to similar studies on ARIMA models for COVID-19 conducted in other countries [6–8]. Thirdly, we used several independent covariates which provided more accurate signals to develop short-term model predictions for immediate outbreak response. And finally, we also optimized the model training and validation period to provide the highest number of data points to generate the best fit model. Potential limitation on the model prediction would include the inherent case data variability, the limited number of data points and errors in data collection. However, data variability was addressed by the use of ARIMA models, as these models take into account data stochasticity in model development is accounted for. In addition, the limited number of data points was address by adding independent covariates and meeting the minimum required data points for ARIMA model development. Furthermore, errors in data collection and diagnostic precision which may adversely affect accuracy of predictions however all COVID-19 case and laboratory data used for this study were verified and validated by the health authorities using the relevant Standard Operating Procedures.

Conclusion

This study demonstrated the effectiveness of ARIMA models as an early warning strategy that can provide accurate COVID-19 forecasts despite limited data points. ARIMA models are not only effective but it's a simple and easy method by which COVID-19 trends can be predicted based on open access data. In addition, the use of smoothened data and independent covariates improved the model accuracy. We are confident that the ARIMA model can be used to generate accurate and reliable forecasts of daily COVID-19 cases beyond 1 May 2020 with the addition of new data points and independent covariates.

Acknowledgements

We would like to thank the Director General of Health Malaysia for his permission to publish this article.

References

1. Gan WH, Lim JW, David KO (2020) Preventing intra-hospital infection and transmission of COVID-19 in healthcare workers. Saf and Health at Work 11: 241-243.
2. Onder G, Rezza G, Brusaferro S (2020) Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. JAMA 323:1775-1776.
3. Hamzah FB, Lau CH, Nazri H, Ligot DV, Lee G, Tan CL (2020) CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ 1: 1-32.
4. Allard R (1998) Use of time-series analysis in infectious disease surveillance. Bull World Health Organ 76: 327–333.
5. Imai C, Hashizume M (2015) A systematic review of methodology: Time series regression analysis for environmental factors and infectious diseases. Trop Med Health 43: 1–9.
6. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in Brief 29: 1-4.
7. Dehesh T, Mardani-Fard HA, Dehesh P (2020) Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. Preprints. 10.1101/2020.03.13.20035345.
8. Chintalapudi N, Battineni G, Amenta F (2020) 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 53: 396-403.
9. Helfenstein U (1996) Box-Jenkins modelling in medical research. Stat Methods Med Res 5:3–22.
10. Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, Lu Z (2013) A hybrid seasonal prediction model for tuberculosis incidence in China. BMC Med Inform Decis Mak 13: 56.
11. Cheung YW, La KS (1995) Lag order and critical values of the augmented dickey-fuller test. J Bus Econ Stat 13: 277-280.
12. Beaumont C, Makridakis S, Wheelwright SC, McGee VE (1984) Forecasting: Methods and applications. J Oper Res Soc 35: 79-81.
13. Balvinder SG (2016) Epidemiology of Dengue in Malaysia from 2005 - 2010 and factors contributing to its emergence. Doctoral thesis. University of Western Australia.93-97 p.
14. Remuzzi A, Remuzzi G (2020) COVID-19 and Italy: what next?. The Lancet 2020 395: 1225- 1228.
15. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLoS One 15: e0231236.
16. Meyler A, Kenny G, Quinn T (1998) Forecasting Irish inflation using ARIMA models. Cent Bank Financ Serv Auth Irel Tech Pap Ser 98:1–49.
17. Barba L, Rodríguez N, Montt C (2014) Smoothing strategies combined with ARIMA and neural networks to improve the forecasting of traffic accidents. Sci World J 2014: 1–14.
18. Kutner MH, Nachtsheim CH, Neter J, Li W (2004) Applied Linear Statistical Models, 5th edition. Boston: McGraw-Hill Irwin Press 1250 p.

Corresponding author

Sarbhan Singh, MBBS. OHD. MPH. DrPH.
Institute for Medical Research
Jalan Pahang, 50588 Kuala Lumpur, Malaysia
Tel: 603-2616 2666
Fax: 603-2693 9335
Email: lssarbhan@moh.gov.my

Conflict of interests: No conflict of interests is declared.

Annex – Supplementary Items

Supplementary Figure 1. Comparison of COVID-19 observed cases, 3 point and 5 point moving average, Malaysia.