MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Wan Amir and Misro | Malaysian Journal of Fundamental and Applied Sciences, Vol. 20 (2024) 1478-1492

RESEARCH ARTICLE

Improving Covid-19 Forecasts in Malaysia: A Hybrid SARIMAX-SARIMA Model with Application to State Elections and Cultural Festivals

Wan Anis Farhah Wan Amir, Md Yushalify Misro*

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Gelugor, Pulau Pinang, Malaysia

Abstract Since the onset of the Covid-19 pandemic, numerous challenges have emerged, including ensuring an adequate supply of personal protective equipment, evaluating the sufficiency of the healthcare workforce, and determining safety measures to sustain businesses and the economy. Consequently, there is a critical need for a computationally competent and realistic model to monitor current caseloads and forecast future cases, thereby enhancing public health awareness, preparation, and response. However, many forecast models currently in use have wide prediction intervals, diminishing their effectiveness as forecasting tools. Thus, this study aims to analyse the trend of Covid-19 cases in Malaysia and develop a forecast model that provides appropriate limits to improve prediction accuracy. This study relied on secondary data of daily Covid-19 cases in Malaysia provided by Ministry of Health from April 12, 2021, to April 24, 2022. Future Covid-19 incidence was predicted using simple, double and Holts-Winter exponential smoothing and SARIMAX models. SARIMAX (0, 1, 1) $(1, 0, 2)^7$ was identified as the best model, exhibiting the lowest error values for forecasting cases. However, the results indicated that SARIMAX's prediction intervals were broad. To address this issue, a new model called hybrid SARIMAX-SARIMA was proposed where the orders from the best SARIMAX model found by using `auto.arima()` function are extracted and used to specify the order for a SARIMA model. The resulting combined model was then utilized to predict future trends in daily Covid-19 cases and evaluation during cultural festivals and state elections. It was observed that the proposed model outperformed others, demonstrating lower error rates and narrower confidence intervals for future predictions.

Keywords: Covid-19, time series forecasting, SARIMAX model, confidence intervals.

***For correspondence:**
yushalify@usm.my

Received: 27 May 2024
Accepted: 18 Nov. 2024

Introduction

Since the discovery of a cluster of unusual pneumonia cases in Wuhan, China, in late December 2019, the world has grappled with the Covid-19 pandemic, the fifth global pandemic since the 2009 H1N1 flu. The rapid spread of the virus led the World Health Organization to declare it a pandemic in March 2020, as reported by Liu, Kuo, and Shih [1]. By April 2022, WHO had confirmed over 505 million cases and more than six million deaths worldwide, with few areas remaining unaffected. Although the outbreak began in China, the majority of cases and fatalities occurred outside the country, particularly in Asia, with Latin America, South America, and the Caribbean experiencing some of the highest death rates globally. The United States was also severely impacted, reporting 80.2 million cases and 982,551 deaths by CNN [2]. In Malaysia, the number of daily new infections decreased by 44% by April 2022, yet the country still reported nearly 16,000 new cases per day, roughly half the number seen on March 9. On April 1, 2022, Malaysia reopened its borders to fully vaccinated international tourists as part of its transition to the Endemic Phase, as highlighted by Flanders Trade [3]. However, according to Zhao *et al*. [4], several challenges persist, including ensuring an adequate supply of personal protective equipment, evaluating

the sufficiency of healthcare workers and resources, and striking a balance between enforcing safety measures and maintaining economic stability.

In India, Tandon *et al*. [5] conducted research on predicting Covid-19 cases in India and other highly affected countries. They found that the ARIMA (2, 2, 2) model had the lowest error measures, making it the most accurate for forecasting future occurrences. Similarly, Chakraborty and Ghosh [6] developed a hybrid forecasting technique combining ARIMA and Wavelet-based models, which accurately predicted short-term projections of future Covid-19 cases. The hybrid ARIMA-WBF model outperformed individual models in Canada, France, and the United Kingdom. Al-Turaiki *et al*. [7] demonstrated time-series models for forecasting Covid-19 cases, recoveries, and fatalities in Saudi Arabia, with the ARIMA model outperforming the cubic spline model for predicting recoveries and deaths.

In Malaysia, Singh *et al*. [8] developed a prediction model for daily confirmed Covid-19 cases using ARIMA, which accurately estimated daily cases from April 18 to May 1, 2020. Additionally, Tan *et al*. [9] used SARIMA models to forecast Covid-19 case trends, with the best SARIMA model indicating a decrease in cases over the forecasted period, aligning with actual observations within a 25% margin of error. Purwandari *et al*. [11] forecasted confirmed, recovered, and fatal Covid-19 cases in Malaysia using a neural network. While the Multi-Layer Perceptron (MLP) model performed best for predicting cases seven steps ahead, the Extreme Learning Machine outperformed MLP based on test results.

In forecasting, prediction intervals play a crucial role in indicating the uncertainty around point forecasts. Preferably, these intervals should be narrow during stable periods and wider when the underlying data is more volatile as mentioned by Christoffersen [11]. Analysts often perceive wide prediction intervals as a sign of model failure, but in cases of non-stationary data, they may actually be more appropriate, as they better capture increased uncertainty, meaning that narrower intervals are not always preferable if they fail to reflect the true level of uncertainty as noted by Chatfield [12]. The significant difference in prediction intervals widths highlight the importance of proper model identification. Given the difficulty of identifying the true model, this research considers using a hybrid of models to improve forecast reliability by proposing a model that has narrower prediction intervals that appropriately balance model uncertainty with reliable forecasts error.

Given the trends in virus transmission, it is crucial to develop accurate models to monitor present caseloads and forecast future cases, enhancing public health awareness, planning, and response efforts. Statistical prediction methods have been suggested as effective tools for predicting and managing pandemic risks, as proposed by Tandon *et al*. [6]. This study aims to construct an improved forecast model for predicting future Covid-19 cases, employing SARIMAX and SARIMA models to predict Covid-19 illness occurrence. Zhang *et al*. [13] concluded that ARIMA models have demonstrated superior prediction of natural disasters compared to support vector machine and weightless neural network models. The proposed model seeks to refine forecast intervals, thereby improving forecast values.

Methods

This section details the data utilized in the study, trend analysis, and several forecasting methods, including exponential smoothing and the ARIMA model. The ARIMA model may be adjusted to incorporate seasonality, such as with the SARIMA model and SARIMA with an additional variable known as SARIMAX. The evaluation of the models is then presented based on RMSE and MAPE, along with the prediction horizon.

Data Description

The study utilized secondary daily data from the Ministry of Health Malaysia (MOH) open data website. A total of 378 daily Covid-19 cases data points were gathered from April 12, 2021, to April 24, 2022, starting from the implementation of the immunization plan in Malaysia, as reported by Homage [14] and ending during the early phase of the “Transition to Endemic” stage, according to The Edge [15]. Subsequently, 4-week daily cases were forecasted, beginning April 25, 2022, and a 1-week comparison between actual and projected values was conducted for Malaysia's top 3 festivals, namely Eid al-Fitr, Deepavali, and Chinese New Year, as well as the preceding Sabah, Malacca, and Johor state elections.

Trend Analysis

A time series plot was utilized to assess the trend in Covid-19 occurrences. This plot displays data collected over time and can identify trends and patterns in the data. Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were also used to examine the current pattern of daily Covid-19 cases in Malaysia.

Exponential Smoothing

Exponential smoothing, introduced by Brown, Holt, and Winters in the late 1950s [16], has been instrumental in the development of robust forecasting methods, as highlighted by Bastos [17]. This approach computes weighted averages of past data, with weights decreasing exponentially as observations become older, as explained by Hyndman and Athanasopoulos [18]. As a result, more recent observations are given greater weight, enabling exponential smoothing to produce accurate forecasts quickly for a wide range of time series data.

Among the various exponential smoothing methods, Simple Exponential Smoothing (SES) is the most basic. It is ideal for data without significant trends or seasonal patterns and relies on a single parameter, the smoothing constant (α), to generate fitted and forecasted values. To address trends in time series data, Holt's method, or Double Exponential Smoothing (DES), extends SES by incorporating a trend component. This method was further extended to account for seasonality, resulting in the Holt-Winters method. The Holt-Winters method includes forecast and smoothing equations for the level, trend, and seasonal components, each with its own smoothing parameters (α, β, and γ).

Choosing appropriate values for the smoothing parameters is crucial to minimize forecasting errors. These values typically range from 0.1 to 0.5, depending on the stability of the underlying data. Lower values are preferred for stable series, while higher values are used for more volatile series, as noted by Ostertagová and Ostertag [19]. In this study, the forecast package in R was employed to fit the data using exponential smoothing methods. These methods also served as a benchmark for comparison against other forecasting models, as demonstrated in the research by Arikan *et al*. [20].

Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA model predicts time series differently as it aims to explain data autocorrelations, whereas exponential smoothing approaches explain data trends and seasonality, Jain and Mallick [21]. The ARIMA model combines the autoregression (AR) and integrated (I) models, as well as the moving average (MA) model. Based on the prior behaviour of the pattern of daily Covid-19 cases, the autoregressive (AR) model forecasted future behaviour. An autoregressive model of order p is represented by an AR (p) model equation. The "Integrated" (I) term indicates that d nonseasonal changes are required to create a stationary time series. Lastly, future values were predicted using the moving average (MA) model function. The time series pattern will also change if the moving average parameters change. The MA (q) model equation refers to a moving average model of order q. When the data is non-stationary and has no seasonal effect, one way to make predictions is with the help of the ARIMA model. In general, ARIMA *(p, d, q)* model equations may be written as in Eq. 1 below as discussed by Ulyah, Mardianto, and Sediono [22]:

$$\phi_p(\beta)(1-\beta)^d Y_t = \theta_0 + \theta_q(B)a_t \quad (1)$$

where (p, d, q) is the AR order (p), differencing order (d), and MA order (q), $\phi_p(\beta)$ is the coefficient of the nonseasonal AR model with order p, $\theta_q(B)$ is the coefficient of nonseasonal MA model with order q and a_t is the residual at t.

Seasonal ARIMA (SARIMA) Model

The ARIMA model is a versatile linear time series model that may simulate various seasonal and nonseasonal time series, Azadeh *et al*. [23]. A seasonal model is created by adding seasonal terms to ARIMA models, and seasonal data can be modelled. Additional differencing for series having a seasonal component is required to eliminate the seasonality effects. The Seasonal ARIMA or SARIMA model can be written as ARIMA $(P, D, Q)^S$ and is presented by Eq. 2 below:

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D Y_t = \theta_q(B)\Theta_Q(B^S)a_t \quad (2)$$

where (p, d, q) is the AR order (p), differencing order (d) and MA order (q), $(P, D, Q)^S$ is the AR order (P), differencing order (D) and MA order (Q), seasonal order (S) for seasonal data, $\phi_p(\beta)$ is the coefficient of nonseasonal AR model with order p, $\theta_q(B)$ is the coefficient of nonseasonal MA model with order q, $\Phi_P(B^S)$ is the coefficient of seasonal AR (S) with order P, $\Theta_Q(B^S)$ is the coefficient of seasonal MA (S) with order Q and a_t is the residual at t.

Regression with SARIMA Errors (SARIMAX) Model

SARIMAX is a SARIMA model with extra variables, Box and Jenkins [24]. The daily Covid-19 case data for the state of Selangor was utilized as an exogenous variable in the model identification process where the regression model is fitted to the exogenous variable with SARIMA errors. This variable was also used to select the optimal model for predicting future changes of Malaysia Covid-19 daily cases in a study by Tan *et al*. [11]. The exogenous variable can be passed to `auto.arima()` via the `xreg` argument, including the variable requires to be specified. This feature will determine the best ARIMA model to fit the data error. Differentiation is applied to all variables during the estimation process, if necessary, but the final model is still expressed in terms of the original variables. Once the model is complete, the best predictors are selected using the Corrected Akaike's Information Criterion (AICc). Each subset of predictors must undergo the method, and the model with the lowest AICc value is chosen as discussed by Hyndman and Athanasopoulos [17].

Proposed Hybrid SARIMAX-SARIMA Model

Given the wide prediction intervals observed in previous forecasts using SARIMAX model, this study aims to reduce the forecast limits and provide more accurate estimates. As noted by Lee [25], narrowing the prediction intervals helps prevent them from becoming excessively large or unbounded. To improve the prediction intervals, this study adopts a sequential hybrid modelling approach, where the `auto.arima()` function is initially used to identify the best SARIMAX model. The optimal model's autoregressive (AR), differencing (*d*), and moving average (MA) orders are then applied to define the corresponding order for a SARIMA model. By combining the seasonal and trend components of SARIMA with the exogenous variables from SARIMAX, the hybrid model is able to deliver more reliable forecasts.

The `auto.arima()` function in R is based on the Hyndman-Khandakar algorithm, Hyndman and Khandakar [26]. This algorithm combines unit root tests, AICc minimization, and MLE to identify the best SARIMAX model. First, the number of differences (*d*) is determined using KPSS tests. Then, values for *p* and *q* are selected by minimizing AICc after differencing the data *d* times. A stepwise search is used to test model variations, starting with initial models. The model with the lowest AICc is chosen, and variations are tested by adjusting *p* and/or *q* by ±1 or including/excluding the constant. This process continues until no further AICc improvements are found.

After identifying the best model using the `auto.arima()` function, the orders of this model are passed to the `Arima()` function for manually fitting SARIMA model. The residuals are then examined by plotting their autocorrelation function (ACF) and conducting a portmanteau test. Once the residuals are confirmed to resemble white noise, forecasting can be proceeded. The steps for the proposed model after data loading and preprocessing with R programming are shown in Figure 1.

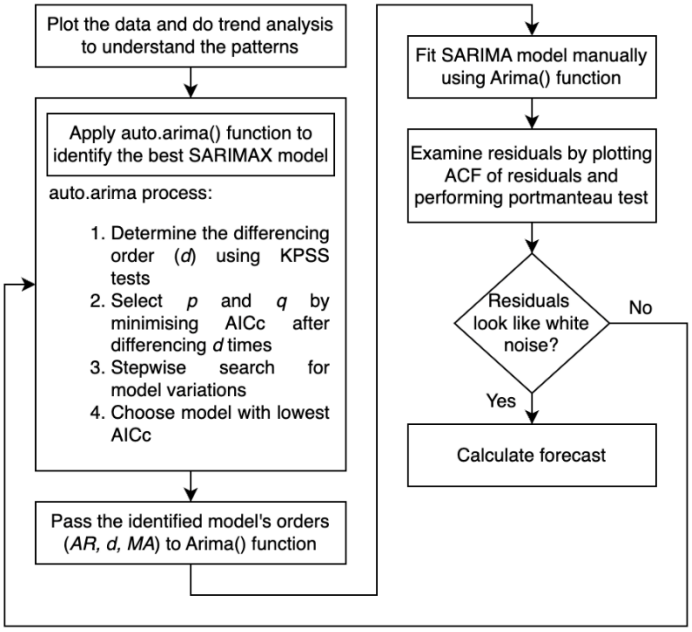

Figure 1. Procedure of the proposed hybrid SARIMAX-SARIMA model

Model Evaluation

The best way to assess a model's predictive performance is by testing it on a separate dataset that is not used during training process, Hyndman and Athanasopoulos [17]. It is common practice to split available data into training data and test data when choosing models, with training data used to estimate the parameters of a forecasting technique and test data used to assess the approach's efficacy. Tiwari [27] has claimed that, depending on the available time series length, the splitting procedure may be accomplished by selecting a split point of the time series in a ratio of 60:40, 70:30, or 80:20. This study employed 378 data points, which was higher than the suggested minimum of 50 data points for providing reliable forecasts, Singh *et al*. [10]. About 80% of the time series data was used for training, and the remaining 20% was used for testing in this study, as shown in Figure 2.

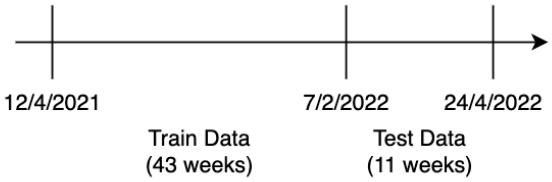

Figure 2. Data split

The test data should indicate the model's performance on new data since it is not used in training. Depending on the length of the sample and how far into the future data must be forecasted, the percentage of the sample used as a test set might vary but is often about 20%. The following error measurements were used to evaluate model performance and choose the optimum model to utilize:

Root mean square error (RMSE)

RMSE is the square root of the mean squared error and taking the root does not affect the relative ranks of models, but it yields a metric with the same units as the, which conveniently represents the typical or "standard" error, Hodson [28]. It is widely used since it is a great all-around error measure for numerical forecasts, Christie and Neill [29]. The formula can be written as Eq. 3 below,

$$RMSE = \sum_{t}^{n} e_t^2 \tag{3}$$

where $e_t = y_t - \hat{y}$, y_t is the actual observed value at time t, and $\hat{y}_t$ is the fitted value at time t.

Mean absolute percentage error (MAPE)

The mean absolute percentage error (MAPE) averages all prediction errors. The MAPE is calculated by adding together all the percentage errors, regardless of their sign. The percentage error provided by this metric makes it simple to interpret, Swamidass, [30]. The formula can be written as Eq. 4 below,

$$MAPE = \sum_{t}^{n} \frac{\left|\frac{e_t}{y_t} \cdot 100\right|}{n} \tag{4}$$

where $e_t = y_t - \hat{y}$, y_t is the actual observed value at time t, and $\hat{y}_t$ is the fitted value at time t.

Forecasting

This research used regular point forecasts instead of the regression model with SARIMA errors, where both the regression and ARIMA parts of the model need to be forecasted together. The forecasting of Covid-19 cases in Malaysia was conducted using the order obtained from the SARIMAX model. Forecasting was done for the following four weeks, from April 25, 2022, to May 22, 2022, which is less than half the time frame of the test data set and is considered short-term forecasting. ARIMA models, according to McCrae *et al*., [31] are very good at predicting short-term forecasts. Furthermore, forecast accuracy will decline if periods are extended as concluded by Toharudin *et al*. [32].

Results and Discussion

Trend Analysis for Daily Covid-19 Cases in Malaysia

Figure 3(a) shows increasing and decreasing trends in daily Covid-19 cases in Malaysia. A trend develops when there is a continuous increase or decrease in the data. It does not need to be linear. A trend is also called "changing direction" when it shifts from growing to falling. There is also cyclical behavior with a length of 26 weeks and significant seasonality within each week since there is always a peak in the middle of each week data cycles when it displays fluctuations with no pattern. According to Hyndman and Athanasopoulos [17], autocorrelations for small lags are strong and positive for trending data since close observations are equally small. ACF values in trended time series tend to be positive and decrease with lag. Autocorrelations are stronger for seasonal delays (at multiples of the seasonal frequency) when data is seasonal. As seen in Figure 3(b), the ACF steadily decreases with an increasing delay because of the trend, whereas the scalloped shape is due to the seasons. PACF plot shows a few significant spikes, the data are clearly non-stationary, with some seasonality. The significant spikes are at lags that are multiples of the 7 with spikes cutting off abruptly after the seasonal lag. There are significant spikes at lag 7 to 8 followed by significant spikes at lags 14 to 15, 21 to 22 for daily data with weekly seasonality.

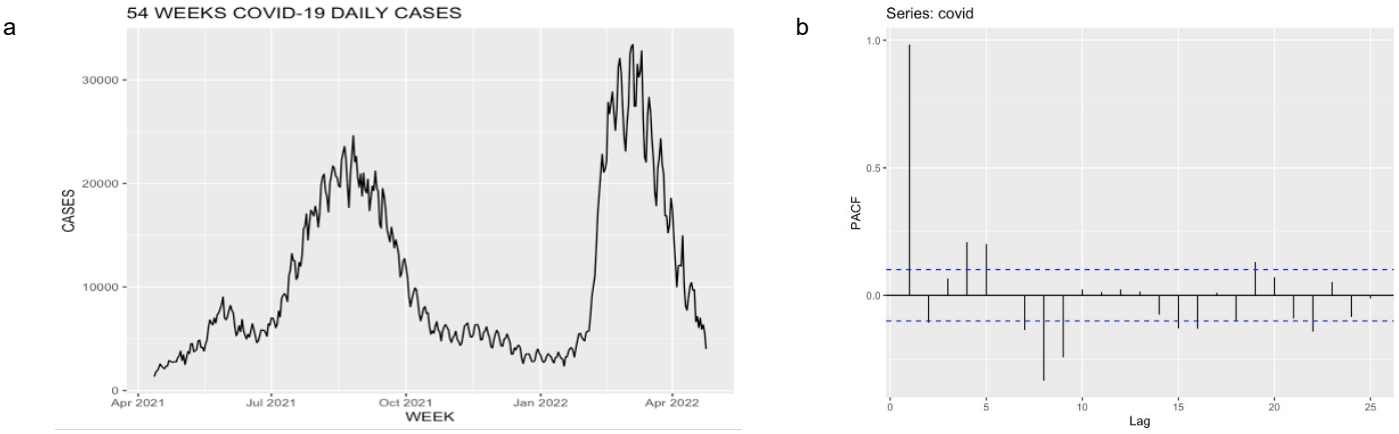

Figure 3. The (a) Covid-19 daily cases in Malaysia and (b) PACF plot for Covid-19 daily cases

Analysis of Exponential Smoothing

This section presents the findings of the exponential smoothing analysis. The findings in Table 1 show that the Holt-Winters method has the lowest RMSE and MAPE scores compared to Single Exponential Smoothing and Double Exponential Smoothing methods. The Holt-Winters method captures seasonality, as the data is affected by the seasonality, this method is unarguably the best model among exponential smoothing methods. A low error values suggest the forecasting model is reliable as proved by Chang, Wang, and Liu [33] and Ishak, Othman, and Harun [34]. This means Holt-Winter can be categorized as an excellent forecasting method.

Table 1. Summary evaluation of exponential smoothing models

Model	Error Measurement	
	RMSE	MAPE
Single exponential smoothing	954.31	9.20
Double exponential smoothing	949.01	9.12
Holt-Winters	769.47	7.47

Analysis of SARIMAX Model

Figure 4(a) shows the time series plot for the train data set, which comprises 43 weeks of data. The time series plot shows that the data is not stationary, as each week has an increasing and decreasing trend and a seasonal component. There is an apparent W-type pattern repeating, so it has seasonality. Non-

stationarity in the data was shown by plotting the ACF and PACF. Figure 4(b) shows a progressive decline in ACF as lags increase in the train data set due to trending, while the "scalloped" shape is seasonal. There are also spikes beyond the blue dotted lines in both plots, indicating that the data are highly autocorrelative. In the ACF plot, the spikes are statistically significant for lags up to 25 indicating a strong correlation in the data over these lags, as explained by Anderson [35]. During the model identification phase, many combinations of the independent variable (the number of daily Covid-19 cases in Selangor) and the dependent variable were tested. The `auto.arima()` function with an `xreg` argument was applied to handle the regression term (independent variable). The function figures out the best combination of *p*, *d*, and *q* parameters using the AIC and BIC (Bayesian Information Criterion) values. Lower values for the AIC and BIC indicate a better fit to the data; hence, these statistics may be used to compare different models as affirmed by Witherspoon *et al*. [36]. The best model generated using this function is shown in Table 2. The portmanteau test returns a high *p*-value for the SARIMAX model, suggesting that the residuals are white noise.

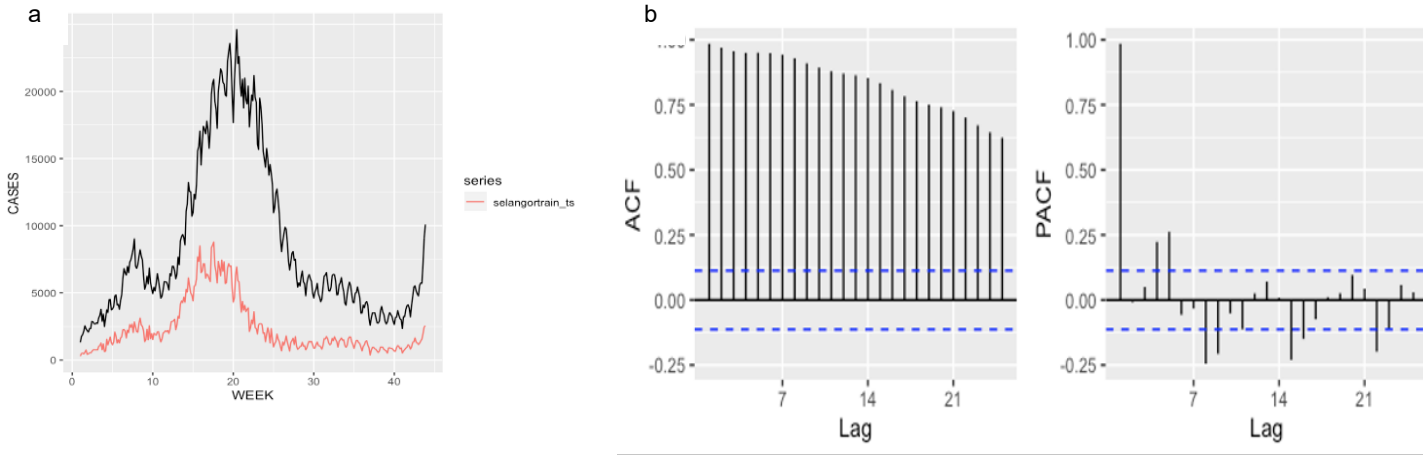

Figure 4. The (a) train data set and (b) ACF and PACF plots for train data set

Table 2. Summary of portmanteau test of the SARIMAX model

SARIMAX (2, 1, 2) (1, 0, 2)^7	
Calculated Q	5.8011
P-value	0.4458
Conclusion	The errors are white noise
RMSE	564.37
MAPE	4.84

Hybrid Model of SARIMAX-SARIMA

To improve the forecast limit of the SARIMAX model, this study proposed a hybrid model by applying the SARIMAX model's parameters to the SARIMA model. Forecasts of Covid-19 cases for the next four weeks were made using the SARIMA model with the parameters arranged in the following order: (2, 1, 2) (1, 0, 2)^7.

The hybrid SARIMAX-SARIMA model's results are summarized in Table 3 and Figure 6. The plots in Figure 5 show that the SARIMA (2, 1, 2) (1, 0, 2)^7 model residuals are normally distributed, and the autocorrelation is close to zero as most spikes are within the bounds on a graph of the ACF (the blue dashed lines) indicating that the residuals are behaving like white noise. A portmanteau test from Table 3 returns a *p*-value more that 0.05, also suggesting that the residuals are white noise.

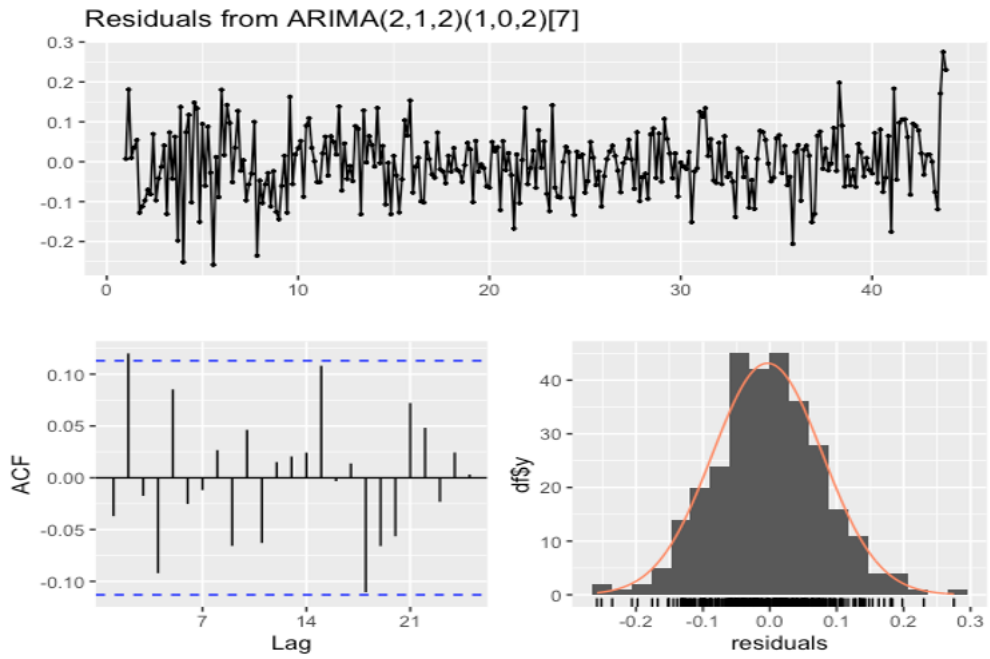

Figure 5. Residual plots of the hybrid SARIMAX-SARIMA model

Table 3. The summary of the hybrid SARIMAX-SARIMA model

Hybrid SARIMAX-SARIMA model	
Calculated Q	13.907
P-value	0.052
Conclusion	The errors are white noise
RMSE	737.17
MAPE	6.53

Evaluation of Model Performance

The best model was selected by comparing its RMSE and MAPE to the other models in Table 4. SARIMAX (2, 1, 2) (1, 0, 2)^7 is found to have the lowest RMSE and MAPE values among all the fitted models followed by the proposed hybrid SARIMAX-SARIMA model. It can also be observed that the RMSE and MAPE values for the test data are slightly higher than those for the training data, however, the difference is minimal and should not be a concern.

Table 4. Summary of train and test data set

Model	Error measurement			
	Train data		Test data	
	RMSE	MAPE	RMSE	MAPE
Single exponential smoothing	954.31	9.20	2474.44	11.54
Double exponential smoothing	949.01	9.12	2461.82	11.63
Holt-Winter	769.47	7.47	1691.20	8.69
SARIMAX (2, 1, 2) (1, 0, 2)^7	564.37	4.84	1698.77	7.46
SARIMAX-SARIMA (2, 1, 2) (1, 0, 2)^7	737.17	6.53	1945.26	8.22

Comparison of standard SARIMA, SARIMAX, and the proposed hybrid SARIMAX-SARIMA models are summarised in Table 5. Based on the table, the result showed that the proposed model is the best option for reaching the goals of low error values and narrow prediction intervals.

Table 5. Comparison of the Box-Jenkins and the proposed models

Model	Error measurement	White noise	Prediction intervals
SARIMA (2, 1, 1) (0, 1, 1)^7	The highest	Has	The narrowest
SARIMAX (2, 1, 2) (1, 0, 2)^7	The lowest	Has	Wide
SARIMAX-SARIMA (2, 1, 2) (1, 0, 2)^7	The second lowest	Has	Narrow

Forecast of Daily Covid-19 Cases in Malaysia

Since the error measures calculated for the SARIMAX (2, 1, 2) (1, 0, 2)^7 model are the smallest compared to the other models, forecasting future values of the Covid-19 daily cases were supposed to be generated using this model. The SARIMAX model requires both regression and ARIMA forecasts to be performed, with the combined results used to make predictions. As shown in Figure 6(a), predictions were obtained by fitting a SARIMAX (2, 1, 2) (1, 0, 2)^7 error model to the percentage change in Malaysian cases with the per cent change in Selangor cases. However, the model's prediction limits are wide; thus, this research proposed a hybrid model by applying SARIMAX model's parameters into a SARIMA model, which is called as SARIMAX-SARIMA model, to narrow these gaps as shown in Figure 6(b). The comparison of the prediction intervals for the SARIMAX and the suggested hybrid SARIMAX-SARIMA models can be seen in Figure 6. The graphs show that the hybrid model's forecast limits are more constrained than standard SARIMAX's.

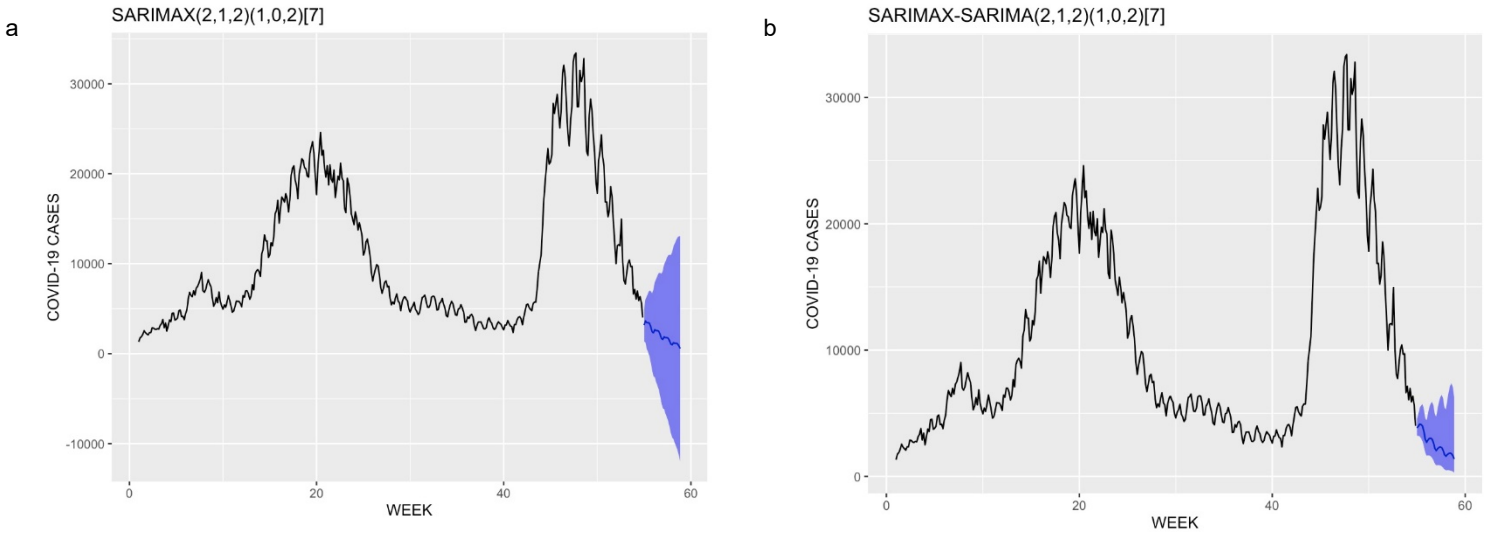

Figure 6. The 4-week Covid-19 forecast by using (a) SARIMAX model and (b) hybrid SARIMAX-SARIMA model

Table 6 displays more information about the predicted data for Covid-19 cases. The actual and predicted values are contrasted, and an absolute percentage error (APE) is presented. Compared to the standard SARIMAX model, the error values produced by the suggested hybrid model are smaller, indicating that the proposed model makes more accurate case forecasts. For the prediction interval, a 95% interval was used in the forecasting; therefore, the actual number of cases is expected to fall within that range in 95% of the time. The cases, however, have a 5% probability of not fitting into this range. Increasing forecast accuracy is one of the preventive measures that can be taken to keep real future value from falling outside the prediction intervals.

Table 6. The comparison of actual and forecast values of Covid-19 cases between SARIMAX and the proposed model

Date	Actual value	SARIMAX model			Hybrid SARIMAX-SARIMA model		
		Forecast value	Error	APE (%)	Forecast value	Error	APE (%)
04/25/22	2478	3171	693	27.97	3387	909	36.68
04/26/22	3361	3644	283	8.42	3970	609	18.12
04/27/22	3471	3444	27	0.78	4150	679	19.56
04/28/22	2935	3453	518	17.65	4102	1167	39.76
04/29/22	2579	3382	803	31.14	3995	1416	54.91
04/30/22	2107	2975	868	41.20	3576	1469	69.72
05/01/22	1503	2438	935	62.21	2941	1438	95.68
05/02/22	1352	2313	961	71.08	2701	1349	99.78
05/03/22	922	2668	1746	189.37	2885	1963	212.91
05/04/22	1054	2561	1507	142.98	3007	1953	185.29
05/05/22	1278	2559	1281	100.23	3024	1746	136.62
05/06/22	1251	2481	1230	98.32	2943	1692	135.25
05/07/22	1372	2136	764	55.69	2653	1281	93.37
05/08/22	2153	1687	466	21.64	2213	60	2.79
05/09/22	2246	1580	666	29.65	2040	206	9.17
05/10/22	2605	1877	728	27.95	2192	413	15.85
05/11/22	3321	1785	1536	46.25	2297	1024	30.83
05/12/22	3410	1787	1623	47.60	2322	1088	31.91
05/13/22	3029	1720	1309	43.22	2271	758	25.02
05/14/22	2373	1439	934	39.36	2058	315	13.27
05/15/22	2239	1070	1169	52.21	1725	514	22.96
05/16/22	1697	983	714	42.07	1597	100	5.89
05/17/22	1469	1226	243	16.54	1724	255	17.36
05/18/22	2017	1151	866	42.94	1814	203	10.06
05/19/22	2124	1152	972	45.76	1841	283	13.32
05/20/22	2063	1098	965	46.78	1808	255	12.36
05/21/22	2021	867	1154	57.10	1644	377	18.65
05/22/22	1817	564	1253	68.96	1384	433	23.83

Comparison of Actual and Forecast Covid-19 Cases in Malaysia During Cultural Festivals and State Elections

Figure 7 and Table 7 illustrate that when compared to the predicted values obtained by the proposed model, none of the actual values for the cultural festivals are entirely bounded within the upper and lower ranges for the one-week forecast. Since Malaysians like celebrating with one another throughout the year, regardless of the holiday, there is a noticeable increase in cases at these times. In 2015, DOSM reported that 63.1% of Malaysians were Malays, 24.6% were Chinese, and 7.3% were Indians, yet this does not stop the country from celebrating its many cultural festivals. WHO [37] has emphasized that the Covid-19 infection rates rise when big people gather for cultural celebrations.

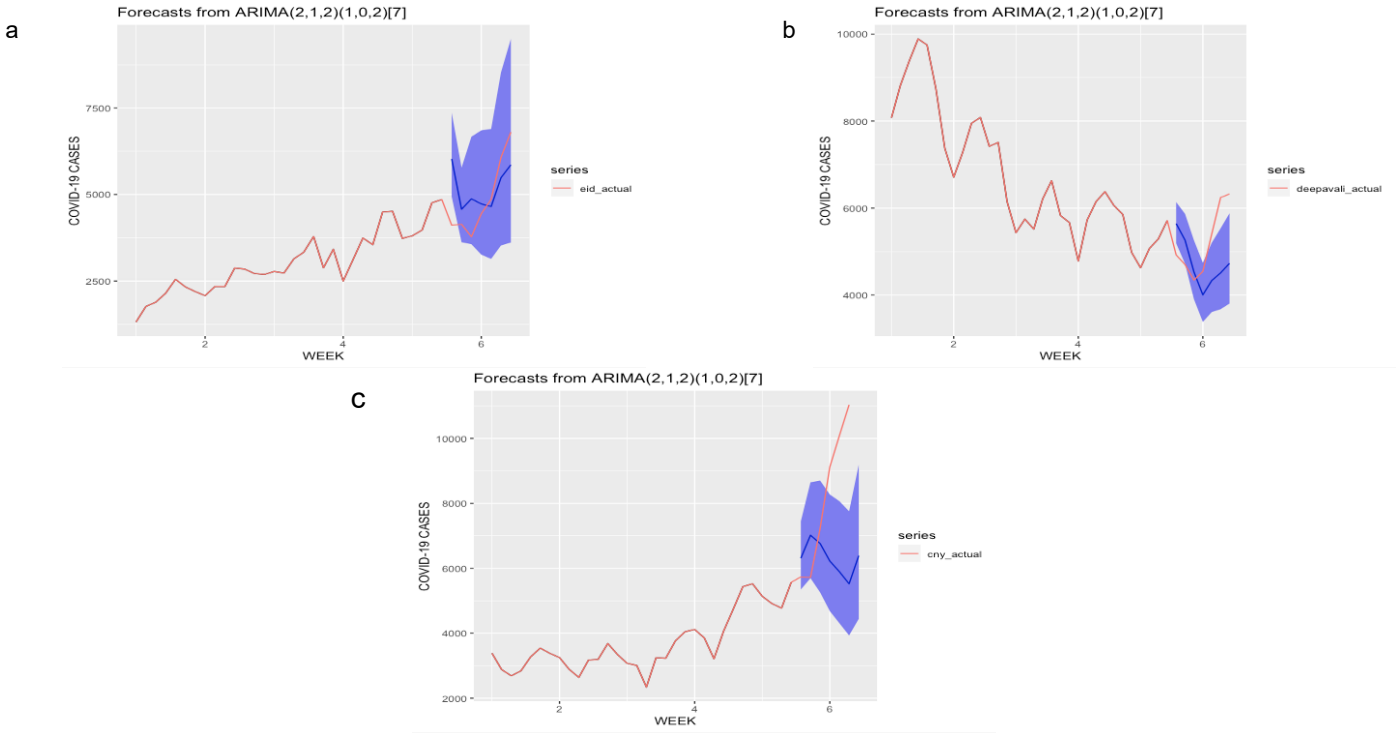

Figure. 7. The comparison of forecast and actual Covid-19 Cases during (a) Eid, (b) Deepavali and (c) Chinese New Year cultural festivals

Table 7. Forecast and actual values of daily Covid-19 cases 1-week after Malaysia cultural festivals

Eid al-Fitr festival				
Date	Forecast value	Lower limit	Upper limit	Actual value
14/05/2021	6 031	4 940	7 364	4 113
15/05/2021	4 574	3 622	5 776	4 140
16/05/2021	4 877	3 565	6 672	3 780
17/05/2021	4 730	3 263	6 855	4 446
18/05/2021	4 654	3 140	6 897	4 865
19/05/2021	5 487	3 529	8 531	6 075
20/05/2021	5 859	3 614	9 499	6 806
Deepavali festival				
Date	Forecast value	Lower limit	Upper limit	Actual value
05/11/2021	5 644	5 190	6 137	4 922
06/11/2021	5 262	4 720	5 866	4 701
07/11/2021	4 531	3 904	5 258	4 343
08/11/2021	4 003	3 380	4 740	4 543
09/11/2021	4 332	3 608	5 202	5 403
10/11/2021	4 510	3 674	5 535	6 243
11/11/2021	4 730	3 804	5 881	6 323
Chinese New Year festival				
Date	Forecast value	Lower limit	Upper limit	Actual value
01/02/2022	6 311	5 348	7 448	5 566
02/02/2022	6 763	5 257	8 700	5 720
03/02/2022	6 763	5 257	8 700	5 720
04/02/2022	6 224	4 683	8 272	7 234
05/02/2022	5 893	4 306	8 065	9 117
06/02/2022	5 522	3 932	7 754	10 089
07/02/2022	6 390	4 444	9 189	11 034

In the three months preceding the Sabah elections, Malaysia only recorded 16 verified cases daily with single-digit or double-digit increases. However, as seen in Figure 8(a) and Table 8, coronavirus infections

rose in Malaysia when thousands of political campaigners and cabinet officials came home without observing the 14-day mandated quarantine. It caused 70% of Covid-19 cases in Sabah after the election and 64.4% of cases in the rest of Malaysia as stated by Lim *et al*. [38]. Despite the unexpected increase in the number of cases, the forecast model can still predict the cases reasonably well, as the actual values are still within the bounds of the prediction. Figure 8(b) and 8(c) along with their respective Table 8 demonstrate that the actual cases of Covid-19 after the Melaka and Johor state elections are within the upper and lower limits and are close to the forecast cases. This means that the forecast model is successful in accurately forecasting Covid-19 cases.

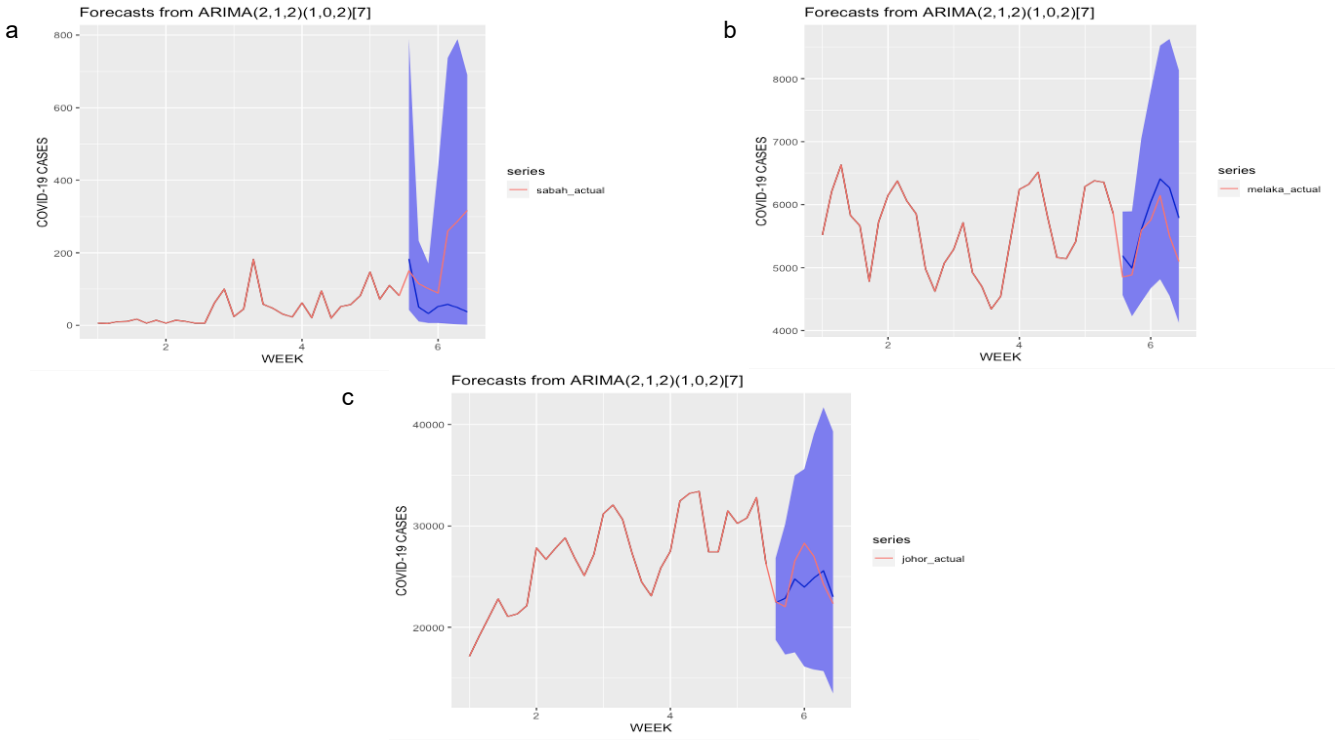

Figure. 8. The comparison of forecast and actual Covid-19 Cases during (a) Sabah, (b) Melaka and (c) Johor state elections

Table 8. Forecast and actual values of daily Covid-19 cases 1-week after Malaysia state elections

Sabah				
Date	Forecast value	Lower limit	Upper limit	Actual value
27/09/2020	183	43	788	150
28/09/2020	50	11	233	115
29/09/2020	33	6	171	101
30/09/2020	52	6	434	89
01/10/2020	58	5	738	260
02/10/2020	49	3	789	287
03/10/2020	37	2	690	317
Melaka				
Date	Forecast value	Lower limit	Upper limit	Actual value
21/11/2021	5 186	4 565	5 891	4 854
22/11/2021	4 993	4 229	5 895	4 885
23/11/2021	5 606	4 455	7 053	5 594
24/11/2021	6 043	4 673	7 815	5 755
25/11/2021	6 408	4 817	8 526	6 144
26/11/2021	6 269	4 553	8 630	5 501
27/11/2021	5 790	4 124	8 130	5 097

Johor				
Date	Forecast value	Lower limit	Upper limit	Actual value
13/03/2022	22 435	18 762	26 826	22 535
14/03/2022	22 857	17 305	30 191	22 030
15/03/2022	24 765	17 535	34 978	26 534
16/03/2022	23 966	16 127	35 616	28 298
17/03/2022	24 868	15 831	39 064	27 004
18/03/2022	25 572	15 681	41 703	24 241
19/03/2022	23 016	13 471	39 324	22 341

Conclusions

The goals of this research were to understand the existing pattern of Covid-19 cases in Malaysia, create a robust forecasting model by narrowing the prediction intervals and predict the pattern of Covid-19 cases in the future. From the initial day of observation to August 2021, the daily Covid-19 cases in Malaysia show an upward trend. The trend fell until early February 2022, when the Omicron wave boosted it. Weekly seasonality was also present, with a peak in the middle of each week and 26-week cyclicity. SARIMAX, single and double exponential smoothing, was used to develop a forecasting model. The result showed that SARIMAX (0, 1, 1) (1, 0, 2)^7 gives the lowest RMSE and MAPE values. This study was then combined the Box-Jenkins SARIMAX and SARIMA models by applying SARIMAX's parameter values into SARIMA model to forecast Covid-19 cases in Malaysia over the next four weeks with smaller prediction intervals. The result demonstrated that the hybrid approach provides a more accurate forecast with a narrower prediction interval compared to other models.

According to the findings, the predicted number of cases would decline over the next four weeks. Although this is a positive sign, proper planning, and adherence to Covid-19 Standard Operating Procedures (SOP) must always be followed. Forecasting models benefit the government and medical practitioners because they serve as an early warning sign, allowing them to better plan for future events and have a healthcare system better equipped to monitor diseases by predicting probable pandemic waves. The forecast values can be utilized to raise public awareness because high Covid-19 cases can seriously affect people's safety and health. The government can also develop an appropriate strategy for organizing future Covid-19 case actions in Malaysia by referring to the graph's pattern, which reveals a seasonal trend and many Covid-19 cases, particularly during the holiday season. Researchers should analyse daily Covid-19 cases in future studies using the most recent data and a more extended period. They are also encouraged to use better analytical models to predict future daily cases, which can be used in Malaysia and worldwide.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

This research was funded by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2023/STG06/USM/03/4) and the School of Mathematical Sciences, Universiti Sains Malaysia. The authors are very grateful to the anonymous referees for their valuable suggestions.

References

[1] Liu, Y. C., Kuo, R. L., & Shih, S. R. (2020). COVID-19: The first documented coronavirus pandemic in history. *Biomedical Journal, 43*(4), 328–333. https://doi.org/10.1016/j.bj.2020.04.007

[2] Tracking COVID-19's global spread. (2022, April 3). *CNN*. Accessed April 20, 2022. https://edition.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/

[3] CORONAVIRUS – The situation in Malaysia | Flanders Trade. (2022, August 10). *Flanders Trade*. Accessed April 24, 2022. https://www.flandersinvestmentandtrade.com/export/nieuws/coronavirus-%E2%80%93-situation-malaysia

[4] Zhao, H., Merchant, N., McNulty, A., Radcliff, T. A., Côté, M. J., Fischer, R., Sang, H., & Ory, M. G. (2021). COVID-19: Short term prediction model using daily incidence data. *PloS One, 16*(4), e0250110. https://doi.org/10.1371/journal.pone.0250110

[5] Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2022). Coronavirus (COVID-19): ARIMA-based time-series analysis to forecast near future and the effect of school reopening in India. *Journal of Health Management, 24*(3), 373–388. https://doi.org/10.1177/09720634221109087

[6] Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals, 135*(June), 109850. https://doi.org/10.1016/j.chaos.2020.109850

[7] Al-Turaiki, I., Almutlaq, F., Alrasheed, H., & Alballa, N. (2021). Empirical evaluation of alternative time-series models for COVID-19 forecasting in Saudi Arabia. *International Journal of Environmental Research and Public Health, 18*(16), 8660. https://doi.org/10.3390/ijerph18168660

[8] Singh, S., Sundram, B. M., Rajendran, K., Law, K. B., Aris, T., Mohd 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

[9] Tan, C. V., Singh, S., Lai, C. H., Md Zamri, A. S. S., Dass, S. C., Aris, T., Mohd Ibrahim, H., & 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

[10] Purwandari, T., Zahroh, S., Hidayat, Y., Sukonob, S., Mamat, M., & Saputra, J. (2022). Forecasting model of COVID-19 pandemic in Malaysia: An application of time series approach using neural network. *Decision Science Letters, 11*(1), 35–42. https://doi.org/10.5267/j.dsl.2021.10.001

[11] Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review, 39*(4), 841. https://doi.org/10.2307/2527341

[12] Chatfield, C. (2001). Prediction intervals for time-series forecasting. In *International series in management science/operations research* (pp. 475–494). https://doi.org/10.1007/978-0-306-47630-3_21

[13] Zhang, Y., Yang, H., Cui, H., & Chen, Q. (2019). Comparison of the ability of ARIMA, WNN, and SVM models for drought forecasting in the Sanjiang Plain, China. *Natural Resources Research, 29*(2), 1447–1464. https://doi.org/10.1007/s11053-019-09512-6

[14] Your guide to COVID-19 vaccinations in Malaysia - Homage Malaysia. (2022). *Homage*. Accessed November 25, 2022. https://www.homage.com.my/resources/covid-19-vaccine-malaysia/

[15] Malaysia to transition to endemic phase of COVID-19 on April 1, says PM. (2022, March 22). *The Edge Malaysia*. Accessed March 24, 2022. https://theedgemalaysia.com/article/malaysia-enter-endemic-phase-april-1-says-pm

[16] Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw-Hill.

[17] Bastos, J. (2019). Forecasting the capacity of mobile networks. *Telecommunication Systems, 72*(2), 231–242. https://doi.org/10.1007/s11235-019-00556-w

[18] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. Accessed March 24, 2022. OTexts.com/fpp2

[19] Ostertagová, E., & Ostertag, O. (2012). Forecasting using simple exponential smoothing method. *Acta Electrotechnica Et Informatica, 12*(3). https://doi.org/10.2478/v10198-012-0034-2

[20] Arikan, B. B., Jiechen, L., Sabbah, I. I. D., Ewees, A. A., Homsi, R., & Sulaiman, S. O. (2021). Dew point time series forecasting at the North Dakota. *Knowledge-Based Engineering and Sciences, 2*(2), 24–34. https://doi.org/10.51526/kbes.2021.2.2.24-34

[21] Jain, G., & Mallick, B. (2017). A study of time series models ARIMA and ETS. *International Journal of Modern Education and Computer Science, 9*(4), 57–63. https://doi.org/10.5815/ijmecs.2017.04.07

[22] Ulyah, S. M., Mardianto, M. F. F., & Sediono. (2019). Comparing the performance of seasonal ARIMAX model and nonparametric regression model in predicting claim reserve of education insurance. *Journal of Physics: Conference Series, 1397*(1), 012074. https://doi.org/10.1088/1742-6596/1397/1/012074

[23] Azadeh, A., Saberi, M., Gitiforouz, A., & Saberi, Z. (2009). A hybrid simulation-adaptive network-based fuzzy inference system for improvement of electricity consumption estimation. *Expert Systems with Applications, 36*(8), 11108–11117. https://doi.org/10.1016/j.eswa.2009.02.081

[24] Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

[25] Lee, B. H. (2022). Bootstrap prediction intervals of temporal disaggregation. *Stats, 5*(1), 190–202. https://doi.org/10.3390/stats5010013

[26] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *DOAJ*. https://doaj.org/article/9b35f41cb88047e78e3d8edab6cd8d99

[27] Tiwari, A. (2021, December 15). Build evaluation framework for forecast models - Towards Data Science. *Medium*. Accessed April 10, 2022. https://towardsdatascience.com/build-evaluation-framework-for-forecast-models-fbc1bd775edd

[28] Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development, 15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

[29] Christie, D., & Neill, S. P. (2022). Measuring and observing the ocean renewable energy resource. In *Elsevier eBooks* (pp. 149–175). https://doi.org/10.1016/b978-0-12-819727-1.00083-2

[30] Swamidass, P. M. (2000). *Encyclopedia of production and manufacturing management*. Springer Science & Business Media.

[31] McCrae, M., Lin, Y. X., Pavlik, D., & Gulati, C. (2002). Can cointegration-based forecasting outperform univariate models? An application to Asian exchange rates. *Journal of Forecasting, 21*(5), 355–380. https://doi.org/10.1002/for.824

[32] Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Kendogo, P., Sijabat, N., Puspita Sari, M. D., Gio, P. U., Basyuni, M., & Pardamean, B. (2021). National vaccination and local intervention impacts on COVID-19 cases. *Sustainability, 13*(15), 8282. https://doi.org/10.3390/su13158282

[33] Chang, P. C., Wang, Y. W., & Liu, C. H. (2007). The development of a weighted evolving fuzzy neural network for PCB sales forecasting. *Expert Systems with Applications, 32*(1), 86–96.

https://doi.org/10.1016/j.eswa.2005.11.021
[34] Ishak, I., Othman, N. S., & Harun, N. H. (2022). Prediction of the COVID-19 pandemic's impact on economy using ARIMA model. *The Journal of Asian Finance, Economics, and Business, 9*(8), 1355–1363. https://doi.org/10.13106/jafeb.2022.vol9.no8.1355
[35] Anderson, A. (2015). *Statistics for big data for dummies*. John Wiley & Sons.
[36] Witherspoon, D., May, E., McDonald, A., Boggs, S., & Bámaca-Colbert, M. Y. (2019). Parenting within residential neighborhoods: A pluralistic approach with African American and Latino families at the center. In *Advances in child development and behavior* (pp. 235–279). https://doi.org/10.1016/bs.acdb.2019.05.004
[37] World Health Organization (WHO). (2021, December 23). Staying safe over the holiday season. *World Health Organization*. Accessed April 20, 2022. https://www.who.int/news-room/commentaries/detail/staying-safe-over-the-holiday-season
[38] Lim, J. T., Maung, K., Tan, S. T., Ong, S. E., Lim, J. M., Koo, J. R., Sun, H., *et al.* (2021). Estimating direct and spill-over impacts of political elections on COVID-19 transmission using synthetic control methods. *PLOS Computational Biology, 17*(5), e1008959. https://doi.org/10.1371/journal.pcbi.1008959