MethodsX 13 (2024) 103013

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Repeated time-series cross-validation: A new method to improved COVID-19 forecast accuracy in Malaysia

Azlan Abdul Aziz[a,b,*], Marina Yusoff[c,d,e], Wan Fairos Wan Yaacob[f], Zuriani Mustaffa[g]

[a] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Cawangan Perlis, Arau 02600, Perlis, Malaysia
[b] Statistical Analytics, Forecasting & Innovation (SAFI) Research Interest Group, Universiti Teknologi MARA Cawangan Perlis, Arau 02600, Perlis, Malaysia
[c] College of Computing, Informatics and Mathematics, UiTM Selangor, Shah Alam 40450, Selangor, Malaysia
[d] College of Computing, Informatics and Mathematics, Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM) Shah Alam, Shah Alam 40450, Selangor, Malaysia
[e] Faculty Business, Sohar University, Oman
[f] College of Computing, Informatics and Mathematics, UiTM Cawangan Kelantan, Kota Bharu 15050, Kelantan, Malaysia
[g] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan 26600, Pahang, Malaysia

ARTICLE INFO

Method name:
Repeated Time-Series Cross-Validation

Keywords:
Forecast accuracy
Cross-validation
Forecasting model
Time Series Analysis

ABSTRACT

Forecasting COVID-19 cases is challenging, and inaccurate forecast values will lead to poor decision-making by the authorities. Conversely, accurate forecasts aid Malaysian government authorities and agencies (National Security Council, Ministry of Health, Ministry of Finance, Ministry of Education, and Ministry of International Trade and Industry) and financial institutions in formulating action plans, regulations, and legal acts to control COVID-19 spread in the country. Therefore, this study proposes Repeated Time-Series Cross-Validation, a new data-splitting strategy to identify the best forecasting model that is capable of producing the lowest error measures value and a high percentage of forecast accuracy for COVID-19 prediction in Malaysia. Some of the highlights of the proposed method are:

- A total of 21 models, five data partitioning sets, and four error measures to improve the forecast accuracy of daily COVID-19 cases in Malaysia.
- The best model selected produces the lowest error measure value for the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE).
- The average 8-day forecast accuracy is 90.2 %. The lowest and highest forecast accuracy was 83.7 % and 98.7 %.

* Corresponding author at: College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Cawangan Perlis, Arau 02600, Perlis, Malaysia

E-mail address: azlan172@uitm.edu.my (A. Abdul Aziz).

https://doi.org/10.1016/j.mex.2024.103013
Received 12 August 2024; Accepted 21 October 2024
Available online 30 October 2024

Specifications table

Subject area:	Mathematics
More specific subject area:	Statistics, Time Series Analysis
Name of your method:	Repeated Time-Series Cross-Validation
Name and reference of original method:	Original method: Time series cross-validation References: Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. https://otexts.com/fpp3/tscv.html.
Resource availability:	This study used daily Malaysian COVID-19 data from Jan. 25 2020 to 28 February 2022. The open-source data were provided by the MOH GitHub platform (https://github.com/MoH-Malaysia/covid19-public).

Background

Forecasting analysis (occasionally known as time series analysis) is a branch of statistical models for developing suitable methods or models to predict future values. Forecasting is a decision-making tool that can act as a scanning device that captures the signals of future outcomes based on past events or related factors that influence the outcome of an event of interest [1]. Forecasting the COVID-19 pandemic is challenging, as inaccurate forecast values will lead to poor decision-making by the authorities.

In 2020, the Malaysian Institute of Economic Research (MIER) and JP Morgan (an American multinational investment bank and financial service) projected a marked increase in daily and cumulative COVID-19 cases in Malaysia at the end of March 2020 and mid-April 2020. The MIER predicted an increase of 5–11 % per day in daily COVID-19 cases in Malaysia between 24 and 31 March 2020 [2]. Nevertheless, the Ministry of Health Malaysia (MOH) [3] reported a significant gap between the actual data and forecasted values, with a forecast percentage error of 22–85 % (Fig. 1). JP Morgan predicted that Malaysia had entered the “acceleration phase” of increasing daily cases of COVID-19 and that it could peak by mid-April 2020 at approximately 6300 cumulative cases [4]. Nonetheless, only 4987 cases were recorded [5], where the JP Morgan forecast strayed with a percentage error of 26 %. Thus, the MIER and JP Morgan projections were inaccurate.

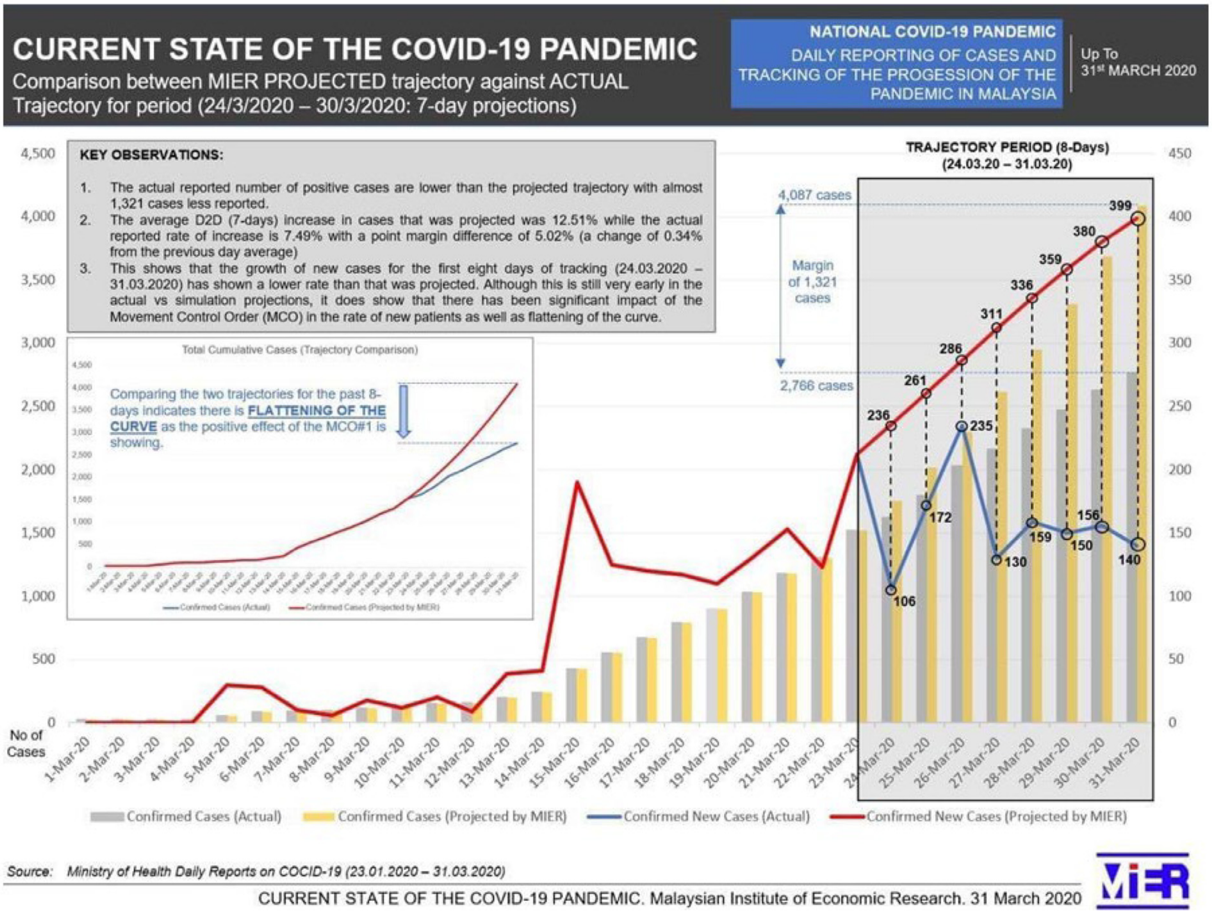

Fig. 1. The MIER projections for future daily COVID-19 cases in Malaysia in 2020.

Inaccurate forecasts would affect Malaysian government authorities and agencies [the National Security Council (MKN), MOH, Ministry of Finance (MOF), Ministry of Education (MOE), Ministry of International Trade and Industry (MITI)] and financial institutions in formulating action plans, regulations, and legal acts to control COVID-19 spread in the country. Accurate forecasting would benefit the MKN, MOH, and MOE and the Ministry of Higher Education (MOHE) in determining regulations and acts, such as the Movement Control Order (MCO), formulating recovery plans and providing necessary medical facilities and staff, and redesigning the learning and teaching process during MCO and resetting the national examination calendar, respectively. Furthermore, accurate forecasting would aid the MOF and MITI in formulating an assistance plan for people affected by COVID-19 and identifying the affected economic sectors and assisting companies, especially small and medium-sized enterprises (SME), respectively.

Policymakers or decision-makers need accurate input to make a decision. Such input can be obtained from the predictive value of the statistical model. The information would enable decision-makers to take the necessary actions to modify plans or to adapt to unexpected environmental changes, such as the COVID-19 pandemic. Therefore, the main objective of this study is to improve the accuracy of predicting COVID-19 in Malaysia by using time series models and a new data-splitting strategy known as Repeated Time Series Cross-Validation (RTS-CV). To achieve the main objective, the following are the sub-objectives that will be studied:

- to determine the best data-splitting ratio for each time series model
- to evaluate the performance of each time series model using statistical criteria.
- to produce forecast values for the next eight days (21–28 February 2022)
- to determine the percentage of forecast accuracy for each forecast value.

Literature review

Data-splitting

Data-splitting, also known as cross-validation, has been used for decades to evaluate the performance of a model. This strategy divides the available data into two parts: one is used to develop a prediction model, and the rest is used to test the model's performance [6]. For example, a study by [7] used the DUPLEX algorithm developed by [8] for data separation to verify the regression model. Their study found that the data-splitting strategy is efficient, but collecting new data to test the model is not practical.

Over time, data-splitting strategies have been widely used in all fields to identify the best model for each studied data set. Some previous studies have shown the success of data splitting, such as the study conducted by [7], where they used the multiple performance validity test (PVT) in the Advanced Clinical Solutions (ACS) package. Their study found that data splitting of these tests can lead to better detection of invalid performance, thus supporting the use of a multifaceted approach in clinical evaluation. The study also concluded that the data-splitting approach can improve the reliability of performance validity assessment in clinical settings. In another study conducted by [8] on smartphone audio recordings from subjects with and without Parkinson's disease, data-splitting by subject is better because it prevents data leakage by ensuring that training and validation are subject-independent. An interesting study conducted by [7] on the Rey Auditory Verbal Learning Test (RAVLT) forced-choice (FC) as a performance validity test (PVT) shows that the use of data-splitting or cross-validation achieves excellent classification accuracy for detecting invalid performance, with a cut-off score ≤ 13 resulting in a sensitivity of 66 % and a specificity of 87 %. Even among patients with normal and mild memory impairment, the FC trial maintained high sensitivity (66 %–82 %) and specificity (≥89 %).

Despite data-splitting or cross-validation is a powerful tool, it has several limitations and weaknesses that can affect its reliability and applicability across different contexts. These limitations include issues related to sample size, sensitivity to outliers, and challenges in specific applications such as time series forecasting. Cross-validation with small sample sizes can affect the reliability and validity of model performance estimates, leading to potential bias and error. [9] his study found that data-splitting with a small sample size can affect the coefficient of determination, R^2 and may not capture the true variability in the data. Meanwhile, another study [10] revealed that the conventional k-fold cross-validation approach can introduce subsampling bias for small datasets. This bias increases the generalization error and reduces the reliability of cross-validation results. Outliers or extreme values can significantly distort cross-validation results and lead to inaccurate estimates of model performance. Outliers can cause overfitting, where a model performs well on training data but poorly on unseen data [11,12]. The temporal dependencies inherent in time series data violate the assumption of independence between observations, a cornerstone of traditional cross-validation methods. Traditional cross-validation methods, such as k-fold cross-validation, assume that data points are independent and identically distributed (i.i.d.). However, time series data are inherently dependent, with each observation potentially influenced by previous ones. This dependency can lead to data leakage, where information from future observations inadvertently influences the model, resulting in overly optimistic performance estimates [13,14]. The effectiveness of cross-validation methods can vary significantly depending on whether the time series is stationary or non-stationary. For stationary time series, blocked cross-validation can be effective, but for non-stationary series, out-of-sample methods like the holdout approach are more reliable [13]. Bayesian models, often used in time series forecasting, face additional challenges with cross-validation due to the computational cost of refitting models multiple times. Approximate methods, such as Pareto smoothed importance sampling, have been developed to mitigate these costs while providing diagnostics on the approximation quality [15].

Method details

Data acquisition

This study used daily Malaysian COVID-19 data from 25 January 2020 to 28 February 2022. The 745 data points included the daily COVID-19 cases in Malaysia. The open-source data were provided by the MOH GitHub platform (https://github.com/MoH-Malaysia/covid19-public). The analysis was conducted using Microsoft Excel, R programming, and RStudio.

Exploration data analysis

The original time series plot examined the data to identify the presence of time series components (trend, seasonal, cyclical, and irregular components), outliers, and missing values.

Model application

This study used 21 time series models. The models included the Naïve model (the benchmark model), the Mean model, State-space models for exponential smoothing (EST) [16], and the Box-Jenkins model [17]. Hyndman et al. [16] and Hyndman and Khandakar [18] introduced the state space models for ETS containing all 18 EST models. This innovation rendered the forecasting process more accessible by automatically generating prediction intervals, likelihood, and model selection criteria based on the model framework. The model framework as shown in Fig. 2.

The Naïve model is most appropriate when historical data have no discernible patterns, such as trends or fluctuations. The Naïve model is usually compared against more complex models. The forecast by the Naïve model is described as follows in Eq. (1):

$$F_{t+m} = y_t \tag{1}$$

where F_{t+m} is the forecast values for m-step ahead made at time t, m refers to the number of step-ahead forecast periods ($m = 1,2,3, \ldots$), and y_t is the last observation at time t.

The mean model assumes the m-step ahead forecast equals the average of all historical data. The general form for the mean model is described in the following Eq. (2):

$$F_{t+m} = \bar{y} \tag{2}$$

Where F_{t+m} is the forecast value for m-step ahead made at time t, m refers to the number of step-ahead forecast periods ($m = 1,2,3, \ldots$), and $\bar{y}$ is the average observation at time t.

The Box-Jenkins model assumes the historical data is stationary and does not exhibit growth or decline over time. The forecast by the ARIMA model is described in Eq. (3) as follows:

$$F_{t+m} = \mu + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{3}$$

where F_{t+m} is the forecast value for m-step ahead made at time t, m refers to the number of step-ahead forecast periods ($m = 1,2,3, \ldots$), μ is the intercept, y_{t-1} is the lagged dependent or current value, and ε_t is the lagged error term.

The Box-Jenkins methodology is generally represented as ARIMA(p,d,q) where p represents the order of the autoregressive part, d represents the degree of differencing involved, and q is the order of the moving average part. Determining p and q values is done by examining the number of significant spikes on the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. However, [18] has created auto.arima function in R programming language to automatically determine the value of $p,d,$ and q. The function combines unit root tests and the Akaike information criterion (AICc) minimization and Maximum Likelihood Estimation (MLE) to obtain an ARIMA model.

Repeated Time Series Cross-Validation (RTS-CV)

The common approach used in univariate time series analysis in assessing model performance to evaluate forecast accuracy is splitting the data into training and testing data. Training data estimates and fits the model parameters, while testing data evaluates its performance. This approach is called cross-validation, a statistical method to assess machine learning model performance. When the study uses high-frequency data such as pandemic outbreak data (COVID-19 or Monkeypox, which is currently hitting several countries in the African continent), the data splitting ratio used can significantly influence the determination of the best model and the accuracy of the forecast value. Typically, the testing data is 20–25 percent of the total data (Fig. 3).

Nevertheless, this approach has drawbacks, such as the forecast accuracy value possibly differing when the testing data size changes [19]. In addition, the main reason for using a 1–5 percent test set size in this study, which is against the conventional approach (20–25 percent), is to ensure minimal data loss for model training. Furthermore, in this study, a test size of 20–25 percent represents a long period of data, which might obscure short-term trends vital for accurate near-term predictions. Using a smaller test set (1–5 percent), the model can be evaluated on more recent observations without sacrificing the long-term training data. Therefore, the optimal forecast accuracy was ensured by introducing the novel RTS-CV data-splitting approach (Fig. 4 and Table 1).

The model performance was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and mean absolute scaled error (MASE).

ADDITIVE ERROR MODELS

Trend	Seasonal N	Seasonal A	Seasonal M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal N	Seasonal A	Seasonal M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathbf{A_d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Fig. 2. State-space equations for each of the models in the ETS framework. Source: https://otexts.com/fpp2/ets.html.

Fig. 3. Example of traditional cross-validation.

Table 1
Training and testing partitioning for each set.

Set		Percentage (%)	Duration	Sample Size
Set 1	Training	99	25 January 2020–20 February 2022	758
	Testing	1	21–28 February 2022	8
Set 2	Training	98	25 January 2020–13 February 2022	751
	Testing	2	14–28 February 2022	15
Set 3	Training	97	25 January 2020–5 February 2022	743
	Testing	3	6–28 February 2022	23
Set 4	Training	96	25 January 2020– 28 January 2022	735
	Testing	4	29 January–28 February 2022	31
Set 5	Training	95	25 January 2020–21 January 2022	728
	Testing	5	22 January–28 February 2022	38

Fig. 4. The RTS-CV.

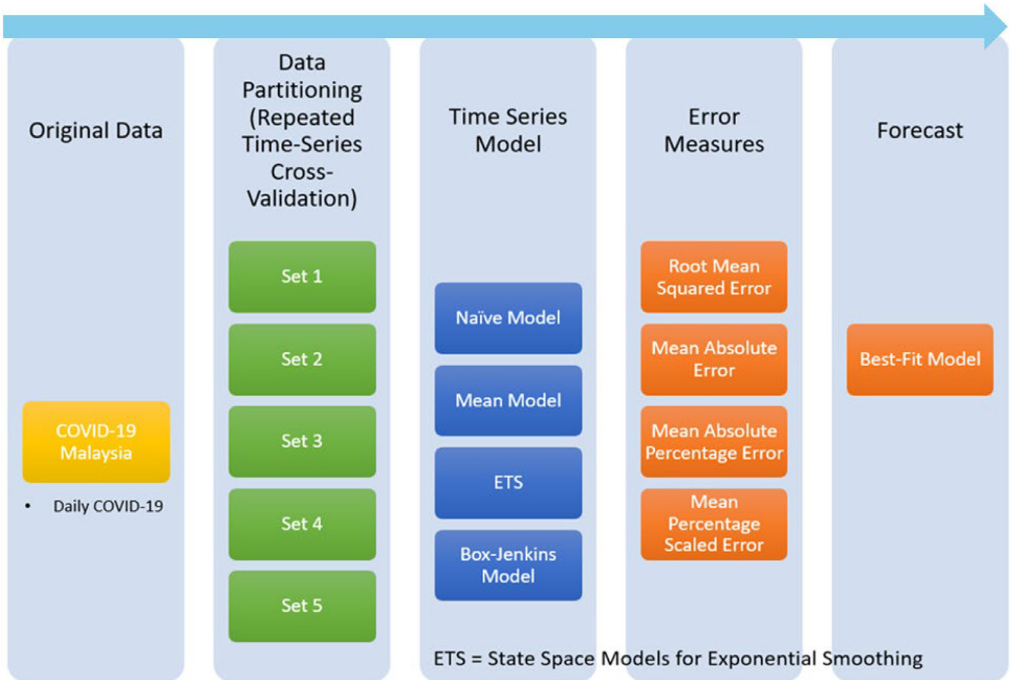

Fig. 5. The forecast model workflow.

Model forecast

Error measures for the training part demonstrate how well the model fits the data. Nonetheless, a model that fits the training part well may not necessarily forecast well [1,20]. Therefore, examining the performance of testing error measures is crucial. The "win" model was selected based on the model that produced the lowest testing error measures. Fig. 4 depicts the workflow of the forecast model.

Method validation

Result analysis

The workflow process (Fig. 5) was demonstrated using daily COVID-19 case data in Malaysia. The dataset was divided into five sets; each had different training and testing percentages. Each model analyzed all five sets, and the training and testing error measures were collected and compared (Table 2). Based on the error measures for each set of partitioning data, the Naïve, ETS, and Box-Jenkins model for Set 1 produced the lowest value for all error measures. In contrast, the mean model produced the lowest error measure in Set 5.

Table 2
Model summary for testing.

	Naïve Model				Mean Model			
	RMSE	MAE	MAPE	MASE	RMSE	MAE	MAPE	MASE
Set 1	3238.56	2757.75	9.81	6.88	23591.42	23381.76	84.42	58.30
Set 2	6533.35	5761.40	20.39	14.82	22992.48	22785.15	84.71	58.61
Set 3	15480.76	14300.61	57.27	38.42	20389.84	19508.95	81.68	52.42
Set 4	16255.22	13519.35	58.43	36.61	17629.20	15043.61	69.14	40.74
Set 5	15803.90	12264.71	58.00	33.30	15932.15	12412.83	59.28	33.70

	ETS Model				Box-Jenkins Model			
	RMSE	MAE	MAPE	MASE	RMSE	MAE	MAPE	MASE
Set 1	3730.75	3174.35	12.09	7.92	3238.56	2757.75	9.81	6.88
Set 2	6533.20	5761.23	20.39	14.82	6705.09	5987.93	21.28	15.40
Set 3	15545.56	14370.74	57.60	38.61	15895.95	14805.15	59.88	39.78
Set 4	16259.98	13524.01	58.45	36.63	16597.30	13855.00	59.94	37.52
Set 5	15816.47	12279.26	58.12	33.34	16099.18	12616.68	61.11	34.26

Smallest error measure for each model.

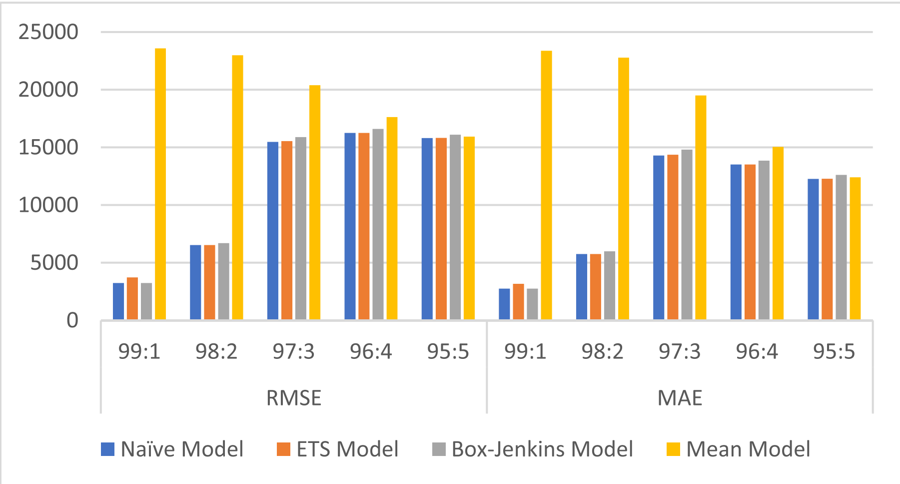

Fig. 6. RMSE and MAE comparison against the size of training and test set.

To illustrate the effectiveness of the proposed data-splitting strategy, Figs. 6 and 7 show that the value of error measures for all models tend to increase following the increase in test size except for the Mean model. Thus, this proves that using a 1–5 percent test set percentage is very efficient in this study in ensuring that the selected model is the best model with the lowest error measures value and a high percentage of prediction accuracy.

Subsequently, the model with the lowest testing error measures was selected and compared to determine the best forecast (Table 3). The best model was then selected from the compared models. Table 3 summarises the models with the lowest error measures.

The Naïve model and Box-Jenkins model produced the lowest error measures. The Naïve model was identical to the ARIMA model (0,1,0); therefore, the naïve or ARIMA model (0,1,0) with Set 1 (training = 99 %, testing = 1 %) was the best model for forecasting daily COVID-19 cases in Malaysia. The predicted value of daily cases of COVID-19 for the next eight days (21–28 February 2022) is expected to be 26,832 cases (Table 4), with an accuracy percentage between 83.7–98.7 percent (mean = 90.2 percent).

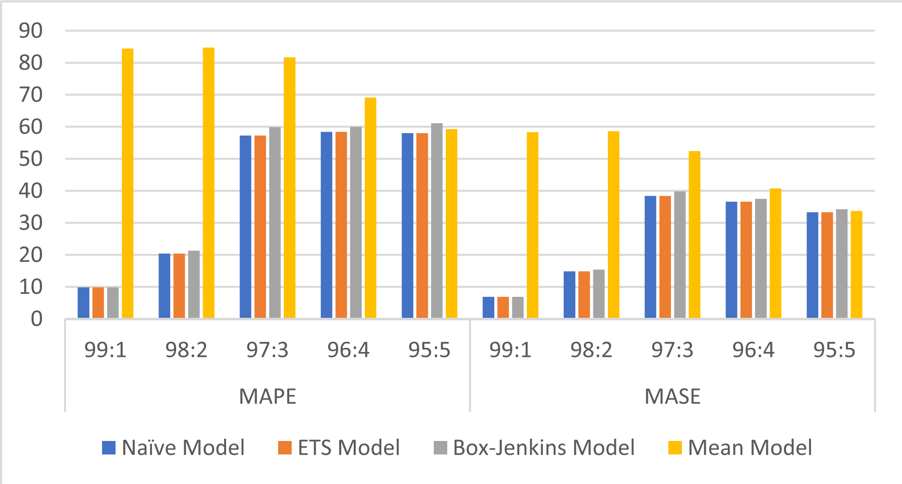

Fig. 7. MAPE and MASE comparison against the size of training and test set.

Table 3
Model comparison for testing.

	Set	Training: Testing (%)	RMSE	MAE	MAPE	MASE
Naïve Model	Set 1	99:1	3238.56	2757.75	9.81	6.88
Mean Model	Set 5	95:5	15932.15	12412.83	59.28	33.70
ETS Model	Set 1	99:1	3730.75	3174.35	12.09	7.92
Box-Jenkins Model ARIMA (0,1,0)	Set 1	99:1	3238.56	2757.75	9.81	6.88

Smallest testing error measures.

Table 4
Eight-step ahead forecast and forecast accuracy.

Date	m	Actual Data	Naïve Model (Set 1)	Accuracy (%)	Range of Accuracy (%)	Average Accuracy (%)
21 February 2022	1	25,099	26,832	93.1	83.7–98.7	90.2
22 February 2022	2	27,179	26,832	98.7		
23 February 2022	3	31,199	26,832	86.0		
24 February 2022	4	32,070	26,832	83.7		
25 February 2022	5	30,644	26,832	87.6		
26 February 2022	6	27,299	26,832	98.3		
27 February 2022	7	24,466	26,832	90.3		
28 February 2022	8	23,100	26,832	83.8		

Discussion

Time series analysis requires more than one data splitting set to enable a more comprehensive comparison of model selection. Subsequently, the best model is selected, yielding a more accurate prediction value. The results suggested that researchers should not focus on only one data-splitting set during model evaluation. A total of 745 data points of daily cases of COVID-19 were analyzed using 21 time series models. This study has successfully met all its objectives. Using a new data-splitting strategy, Repeated Time Series Cross-Validation (RTS-CV), the best data-splitting ratio for this study is 99:1. The higher the test size used, the higher the value of the error measures. The Naive model or ARIMA(0,1,0) is the best after the performance evaluation is carried out with the lowest value of RMSE (3238.56), MAE (2757.75), MAPE (9.81) and MASE (6.88). The forecast for the next eight days of COVID-19 (21–28 February 2022) is estimated at 26,832 cases. In addition, this study found a high percentage of prediction accuracy ranging from 83.7 to 98.7 percent. For future research recommendations, it is recommended that RTS-CV be applied to forecast daily cases of Monkeypox, which is currently affecting several countries on the African continent.

Limitations

A data-splitting ratio of 99:1 may only be suitable for the dataset used in this study. It is also possible that the results will be different by using different datasets. However, the idea behind Repeated Time Series Cross-Validation (RTS-CV) is for researchers to consider several sets of data-splitting so that the forecast value produced is optimal. For example, RTS-CV was used in a study by [21] to find the best model for forecasting the Consumer Price Index in Malaysia. A total of 5 data-splitting sets were used with a test size between 10 and 15 percent. The study results found that a training-test ratio of 80:20 is the best.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Ethics statements

This study used daily Malaysian COVID-19 data from 25 January 2020 to 28 February 2022. The open-source data were provided by the MOH GitHub platform (https://github.com/MoH-Malaysia/covid19-public).

CRediT author statement

Azlan Abdul Aziz: Idea and conceptualization, data curation, methodology, data analysis, and writing the original draft, **Marina Yusoff**: Supervision, review and editing, **Wan Fairos Wan Yaacob**: Supervision, validation, review and editing, **Zuriani Mustaffa**: Supervision, validation, review and editing.

Acknowledgments

The authors express their gratitude to Universiti Teknologi MARA for their invaluable support, and the MOHE Malaysia under the Fundamental Research Grant Scheme (FRGS) (Reference No.: FRGS/1/2022/STG06/UITM/02/1). The authors are also grateful to the late Ustazah Wan Halimatun Saadiyah and Associate Prof. Dr. Puzziawati Abu Ghani for the knowledge and guidance given.

References

[1] R.J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, OTexts (2018).
[2] MIER, “The Economic Impacts of COVID-19,” Mar. 31, 2020.
[3] N.H. Abdullah, Situasi Semasa Jangkitan Penyakit Coronavirus 2019 (COVID-19) di Malaysia, From the Desk of the Director-General of Health Malaysia (2020) https://kpkesihatan.com/2020/04/01/kenyataan-akhbar-kpk. -1-april-2020-situasi-semasa-jangkitan-penyakit-coronavirus-2019-covid-19-di-malaysia/(accessed Apr. 01, 2020).
[4] T. Star, JP Morgan: Covid-19 Likely to Peak Next Month in Malaysia, 2020 Mar. 25,.
[5] T.B.T.M. Adnan, Covid-19: Jangkaan JP Morgan Meleset, Sinar Harian, 2020 Apr. 13,.
[6] R.R. Picard, K.N. Berk, Data Splitting, Am. Stat. 44 (2) (1990) 140–147 May, doi:10.1080/00031305.1990.10475704.
[7] K.M. Bain, et al., Cross-validation of three Advanced Clinical Solutions performance validity tests: examining combinations of measures to maximize classification of invalid performance, Appl. Neuropsychol. Adult 28 (1) (2021) 24–34, doi:10.1080/23279095.2019.1585352.
[8] I. Tougui, A. Jilbab, J. El Mhamdi, Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications, Healthc. Inform. Res. 27 (3) (2021) 189–199 Jul., doi:10.4258/hir.2021.27.3.189.
[9] D. Kovács, P. Király, G. Tóth, Sample-size dependence of validation parameters in linear regression models and in QSAR, SAR QSAR Environ. Res. 32 (4) (2021) 247–268 Apr., doi:10.1080/1062936X.2021.1890208.
[10] L. Guo, J. Liu, R. Lu, Subsampling bias and the best-discrepancy systematic cross validation, Sci. China Math. 64 (1) (2021) 197–210, doi:10.1007/s11425-018-9561-0.
[11] S. Bates, T. Hastie, R. Tibshirani, Cross-Validation: What Does It Estimate and How Well Does It Do It? J. Am. Stat. Assoc. 119 (546) (Apr. 2024) 1434–1445, doi:10.1080/01621459.2023.2197686.
[12] S. Wager, Cross-Validation, Risk Estimation, and Model Selection: Comment on a Paper by Rosset and Tibshirani, J. Am. Stat. Assoc. 115 (529) (Jan. 2020) 157–160, doi:10.1080/01621459.2020.1727235.
[13] V. Cerqueira, L. Torgo, I. Mozeti\vc, Evaluating time series forecasting models: an empirical study on performance estimation methods, Mach. Learn. 109 (2019) 1997–2028 [Online]. Available https://api.semanticscholar.org/ . CorpusID:167217568.
[14] M.M. Malik, A hierarchy of limitations in machine learning, CoRR (2020) abs/2002.0[Online]. Available https://arxiv.org/abs/2002.05193 .
[15] P.-C. Bürkner, J. Gabry, A. Vehtari, Approximate leave-future-out cross-validation for Bayesian time series models, J. Stat. Comput. Simul. 90 (14) (Sep. 2020) 2499–2523, doi:10.1080/00949655.2020.1783262.
[16] R.J. Hyndman, A.B. Koehler, R.D. Snyder, S. Grose, A state space framework for automatic forecasting using exponential smoothing methods, Int. J. Forecast. 18 (3) (2002) 439–454, doi:10.1016/S0169-2070(01)00110-8.
[17] D.J. Bartholomew, G.E.P. Box, G.M. Jenkins, Time series analysis forecasting and control, Oper. Res. Q. 22 (2) (1971), doi:10.2307/3008255.
[18] R.J. Hyndman, Y. Khandakar, Automatic time series forecasting : the forecast package for R Automatic time series forecasting : the forecast package for R, J. Stat. Softw. 27 (3) (2008) 1–22, doi:10.18637/jss.v027.i03.
[19] N.A.A. Haris, A.A. Aziz, N.A.M. Nor, N. Sharif, Improving Air Pollution Index (Api) Predictive Accuracy Using Time Series Cross-Validation Technique, 2018.
[20] M.A. Lazim, Introductory Business Forecasting: A Practical Approach, University Publication Centre (UPENA), 2011.
[21] N.Q. Mohamad Fozi, A.A. Aziz, A novel hybrid holt integrated moving average (hima) for consumer price index prediction, Int. J. Entrep. Manag. Pract. 7 (2024) 25 SE-Articles, Jun.[Online]. Available https://gaexcellence.com/ijemp/article/view/509 .