DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 5, ISSUE 2, 2024, 23 – 27
DOI: https://doi.org/10.15282/daam.v5i2.9705

RESEARCH ARTICLE

Imputation of new COVID-19 cases missing data using basic statistical methods

N.Z.A. Hamid*, A.A. Zambri, N.B.A. Wahid, N.H. Adenan, N.H.M. Husin, N.W.M. Junus, N.S.A Karim and R.A. Tarmizi

Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900, Tanjong Malim, Perak, Malaysia

ABSTRACT - This study investigates the imputation of new COVID-19 cases with missing data in Kedah and Selangor states by using basic statistical methods. This study aims to impute missing data using four basic statistical methods and compare the methods using a performance index. The four basic statistical methods applied in this study are Linear Interpolation, Top Bottom Average, 7-Day Average, and 14-Day Average. The time series data employed is the number of new COVID-19 cases for 365 days, which was recorded in daily numbers for 2021 in Kedah and Selangor states. The time series data was sampled and randomly removed by 10% and 20%. The removed data will be imputed using the four basic statistical methods. The performance indices used to compare the performance of the basic statistical methods were mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (CC). Overall, Linear Interpolation and 14-day Average are suitable basic statistical methods for finding missing data. The findings of this study suggest that basic statistical techniques can be instrumental in supporting the Malaysian Ministry of Health (MOH) in filling gaps in data on new COVID-19 cases in future initiatives.

ARTICLE HISTORY
Received : 5th Sept 2023
Revised : 28th Feb 2024
Accepted : 1st April 2024
Published : 30th Sept 2024

KEYWORDS
Imputation method
Missing data
Basic statistical method
COVID-19 cases

1. INTRODUCTION

According to [1–5] and many more, the COVID-19 pandemic has evolved into a health, socioeconomic and humanitarian crisis of unprecedented scale and impact. The situation in Malaysia is compounded by the fact that the Government came into power only in early March of 2020 and is already facing a heavy debt problem, financial constraints, plummeting oil prices and knock-on effects on trade and tourism from the global shutdown.

On the pandemic front, the Government has received international recognition for its efforts regarding testing, contact tracing, quarantine, and treatment, while keeping first responders safe and providing reliable information and advice to the public. Daily updated information on the numbers and rates of infection, fatalities and recoveries, and identification of 'hot spots', track progress in 'flattening the curve'. A daily number of new COVID-19 cases is recorded to calculate the measure of human populations with infection of the disease. The community is advised to comply with the standard operating procedure (SOP) [6–8] and live their lives conforming to the new norms to ensure that the chain of COVID-19 can be broken. The COVID-19 pandemic has had a huge impact on Malaysian education systems [9] and involves masses of students who could cause a rapid spread of the virus. An investigation carried out by the Ministry of Health (MOH) into the reported cases and clusters of education found that these cases were caused by SOPs non-compliance by the people at the educational institutions during and outside of formal school hours, especially when interacting in the dormitory. Furthermore, according to [10], the manufacturing industry also faces a big impact from COVID-19. Therefore, it is very important to conduct research on COVID-19 cases.

According to [11], missing data frequently occurs in quantitative data analysis. This can result from factors such as recording errors or instrument malfunctions. In the case of COVID-19 new case data, missing values may also arise due to the absence of standardised and systematically collected surveillance data during the COVID-19 outbreak [12]. With the development of computational capabilities, recent methods can be applied to manage the missing data issue. Therefore, this study seeks to impute the missing data of new COVID-19 cases for the states of Kedah and Selangor in 2021 using basic statistical methods of Linear Interpolation [12], Top Bottom Average, 7-Day Average, and 14-Day Average [12], and all were tested using performance indices, namely, mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (CC) to find the most effective method for the process of imputing the missing data.

Linear Interpolation and the 14-day moving average have been employed in prior COVID-19 studies, primarily for trend smoothing and, in some cases, for imputing missing data [12]. In contrast, among the four methods examined, the Top-Bottom Average method has received limited attention in the literature. Nevertheless, it offers a simple and interpretable approach for estimating missing values in the absence of complete COVID-19 new case data. The 7-day average, although widely used for trend monitoring by health authorities such as the Centres for Disease Control and Prevention (CDC), has not, to the best of the authors' knowledge, been applied for imputing missing values in the literature reviewed. This study, therefore, investigates the use of the 7-day average as an additional method to address missing data in COVID-19 new case reports.

***CORRESPONDING AUTHOR | N.Z.A. Hamid**| ✉ nor.zila@fsmt.upsi.edu.my

2. METHODOLOGY

2.1 Time series data of new COVID-19 cases

The time series used in this entire study is new COVID-19 cases collected by day from January 1, 2021, to December 31, 2021. The selected study locations were in the states of Kedah and Selangor. Selangor is Malaysia's most populous state, as well as the state with the largest economy in terms of gross domestic product, while the state of Kedah is mostly rural. They were COVID-19 cases reported in Kedah, for example [3], and Selangor [13]. Thus, both states were selected. The entire duration of the time series is 365 days for each state and was recorded in numerical units.

2.2 Imputation of missing data

The imputation of missing data is applied using time series data of COVID-19 cases from the states of Kedah and Selangor in 2021. To simulate missingness, 10% and 20% of the data were randomly removed. These missingness rates were selected as they were employed in a previous study [12]. While the previous study covered a broader range of missingness rates from 5% to 30%, this study focuses on 10% and 20% to serve as a preliminary investigation in the Malaysian context, aiming to assess the effectiveness of basic statistical methods. Lower missingness rates, such as 5%, may not introduce sufficient variability in the performance of imputation methods, and intermediate rates, like 15%, were excluded to maintain clarity and comparability between the two chosen levels. Missingness rates exceeding 20% are not considered in this study, as such rates typically require more sophisticated imputation methods, such as k-nearest neighbour (KNN), as suggested by Pham et al. [12]. This study, however, is intentionally limited to basic statistical methods. Consequently, the missing data are imputed using four basic statistical methods and evaluated using performance indices to determine the most effective approach.

2.3 Basic statistical methods

Four basic statistical methods were applied to impute missing data that had been randomly removed at rates of 10% and 20%. These methods are Linear Interpolation, Top Bottom Average, 7-Day Average, and 14-Day Average. Most of these methods have been adopted in previous studies [11,14–18], particularly for environmental data such as air pollution relevance data. In contrast, the application of these basic methods to COVID-19-related data remains limited, especially within the Malaysian context. For example, in an international study, Linear Interpolation and 14-Day Average were employed in COVID-19 research [12], while Top Bottom Average and 7-Day Average have not been extensively examined. Since basic statistical methods tend to perform well merely under low missingness conditions, more advanced methods are generally recommended when missingness exceeds 20% [12]. While these methods are widely used, their performance and limitations vary depending on the data and context, with more sophisticated techniques often required for higher missingness rates or more complex data patterns. This preliminary study, therefore, focuses on evaluating the performance of four basic imputation methods on COVID-19 time series data at missingness rates of 10% and 20%, to explore their practical effectiveness in this application domain.

Linear Interpolation (equation 1) combines two data points with a straight line, while the Top Bottom Average (equation 2) is used to compute missing data by averaging above (top) and below (bottom) the missing data. Therefore, missing data can be imputed directly using equations (1)–(2).

$$y^* = y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x^* - x_1) \tag{1}$$

$$y^* = \frac{y_2 + y_1}{2} \tag{2}$$

where y^* is the missing data sought, x^* is the time point of the missing data, x_1 and y_1 are the coordinates above of the missing data, and x_2 and y_2 are the coordinates below of the missing data.

The 7-Day Average (equation 3) is the average observation calculated every 7 days, while the 14-Day Average (equation 4) is the average calculated every 14 days near the missing data.

$$y^* = \frac{\sum_{j=1}^{7} y_j}{7} \tag{3}$$

$$y^* = \frac{\sum_{j=1}^{14} y_j}{14} \tag{4}$$

where j is every 7 days and 14 days, respectively.

2.4 Performance index

In order to determine the best basic statistical method, three performance indices were applied, namely MAE, RMSE, and CC.

$$\text{MAE} = \frac{\sum_{i=1}^{N}|P_i - O_i|}{N} \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}|P_i - O_i|^2}{N}} \tag{6}$$

$$\text{CC} = \frac{\sum_{i=1}^{N}(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(P_i - \bar{P})^2 \sum_{i=1}^{N}(O_i - \bar{O})^2}} \tag{7}$$

where N is the number of imputations, O_i is the observed point, P_i is the imputed data point, $\bar{O}$ is the average of the observed data, and $\bar{P}$ is the average of the imputed data.

3. RESEARCH FINDINGS AND DISCUSSIONS

3.1 Results of the prediction of missing data

The total time series data period for new COVID-19 cases in Kedah and Selangor states in 2021 is 365 days. The time series data that were randomly removed by 10% and 20% were imputed using all four statistical methods and then tested using the performance index to determine the best method. Table 1 and Table 3 show the prediction results of 10% missing data, in which 37 out of 365 data points were randomly removed. Meanwhile, Table 2 and Table 4 demonstrate the prediction results of 20% missing data, where 73 out of 365 data points were randomly removed.

Table 1. Prediction results of 10% missing data of new COVID-19 cases in Kedah in 2021

Basic Statistical Methods	MAE	RMSE	CC
Linear Interpolation	48	74	0.9876
Top Bottom Average	48	74	0.9876
7-Day Average	**34**	**64**	0.9920
14-Day Average	36	67	**0.9931**

Table 2. Prediction results of 20% missing data of new COVID-19 cases in Kedah in 2021

Basic Statistical Methods	MAE	RMSE	CC
Linear Interpolation	88	147	0.9679
Top Bottom Average	90	145	0.9688
7-Day Average	68	110	0.9821
14-Day Average	**66**	**107**	**0.9828**

Table 3. Prediction results of 10% missing data of new COVID-19 cases in Selangor in 2021

Basic Statistical Methods	MAE	RMSE	CC
Linear Interpolation	**262**	**376**	**0.9862**
Top Bottom Average	266	379	0.9860
7-Day Average	394	642	0.9479
14-Day Average	354	596	0.9558

Table 4. Prediction results of 20% missing data of new COVID-19 cases in Selangor in 2021

Basic Statistical Methods	MAE	RMSE	CC
Linear Interpolation	**304**	**450**	**0.9757**
Top Bottom Average	331	498	0.9704
7-Day Average	357	526	0.9679
14-Day Average	328	487	0.9700

3.2 Discussion of the best methods

According to [19], a CC value approaching one indicates that the predicted data closely approximates the actual data. Based on Tables 1 and 2, the 14-Day Average method demonstrates high performance, with CC values of 0.9931 and 0.9828, respectively. Meanwhile, in Tables 3 and 4, the Linear Interpolation method also shows strong performance, with CC values of 0.9862 and 0.9757, respectively. Therefore, among the four basic statistical methods evaluated, the 14-Day Average appears to be the most effective for the state of Kedah, while Linear Interpolation performs best for the state of Selangor. Although the 14-Day Average has been used in previous COVID-19-related studies [12], no prior literature was found that specifically identifies it as the best imputation method. Furthermore, differences in data patterns, reporting practices, and missingness mechanisms across nations limit the comparability of international studies to the current findings, which focus on two Malaysian states.

4. CONCLUSIONS

Basic statistical methods can effectively impute missing data in the COVID-19 case time series for the states of Kedah and Selangor in 2021. Among the four methods evaluated, the 14-Day Average performed best for Kedah, while Linear Interpolation was most suitable for Selangor. This difference in optimal methods likely reflects variations in data characteristics such as reporting frequency, underlying trends, and patterns of missingness between the two states. Therefore, both methods are recommended depending on the regional context. For future research, more advanced imputation techniques, such as KNN, should be considered for comparison. Additionally, testing these methods under different missing data scenarios by varying random seeds can help evaluate their robustness and reliability.

ACKNOWLEDGEMENTS

Instituition(s)
All the authors would like to express their gratitude to Universiti Perguruan Sultan Idris (UPSI) for allowing this study to be conducted. Thank you to the Malaysian Ministry of Health for facilitating the acquisition of data.

Fund
This study was not supported by any grants from funding bodies in the public, private, or non-profit sector.

Individual Assistant
NA

AUTHOR CONTRIBUTIONS

N.Z.A. Hamid (Conceptualization; Formal analysis; Writing- original draft), A.A. Zambri (Conceptualization), N.B.A. Wahid (Formal Analysis); N.H. Adenan (Methodology), N.H.M. Husin (Formal Analysis), N.W.M. Junus (Methodology), N.S.A Karim (Writing- review & editing), R.A. Tarmizi (Writing- review & editing).

DECLARATION OF ORIGINALITY

The authors declare no conflict of interest to report regarding this study.

REFERENCES

[1] Yong SS, Sia JK. COVID-19 and social wellbeing in Malaysia: A case study. Current Psychology. 2023;42(12):9577-91.
[2] Lim LL. The socioeconomic impacts of COVID-19 in Malaysia: Policy review and guidance for protecting the most vulnerable and supporting enterprises. International Labour Organization. 2020:1-99.
[3] Azit NA, Mohd Suan MA, Omar N, Dali N, Romli M, Md Yusof MA, Ahmad M, Ibrahim MZ, Abdul Rahman S. Epidemiological investigation of a covid-19 community cluster in kedah, malaysia. International Journal of Travel Medicine and Global Health. 2022;10(1):10-5.
[4] Hanis TM, Arifin WN, Musa KI, Hasani WS, Nawi CM, Shahrani SA, Chen XW, Suliman MA, Khan EE, Ab Aziz WA, Said MZ. Risk factors for COVID-19 mortality in Malaysia. The Malaysian Journal of Medical Sciences: MJMS. 2022;29(6):123.
[5] Cheng C. Pandemic Economics: the impact of the COVID-19 pandemic on the Malaysian economy. In 2020 RIN Online Workshop Series on COVID-19. Available at: https://d-arch. ide. go. jp/RIN/common/pdf/2020-09_ws-abstract_4-2_calvin. pdf 2020.
[6] Malaysian National Security Council. MySOP. Putrajaya: Majlis Keselamatan Negara. 2022.
[7] Ministry of Health Malaysia. COVID-19 Malaysia. Retrieved from https://data.moh.gov.my/dashboard/covid-19; 2023.
[8] Shafii H, Radzi NA, Yassin AM, Masram H. Implementing COVID-19 Standard Operation Procedure (SOP) in Malaysia Construction Industry: Challenges and Strategies. International Journal of Property Sciences (E-ISSN: 2229-8568). 2022;12(1):37-53.

[9] Sufian SA, Nordin NA, Tauji SS, Nasir MK. The impacts of Covid-19 to the situation of Malaysian education. International Journal of Academic Research in Progressive Education and Development. 2020;9(2):764-74.
[10] Zulfakar HHB, Yusof AMB, Bin Sapian MK, Nallaluthan K. A Review of Covid-19 Pandemic Impacts on Malaysian Manufacturing Industries. Quality and Quantity Research Review. 2021;6(3):116–28.
[11] Ghapor AA, Zubairi YZ, Imon AHMR. Missing value estimation methods for data in linear functional relationship model. Sains Malaysiana. 2017;46(2):317–26.
[12] Pham HT, Do T, Baek J, Nguyen CK, Pham QT, Nguyen HL, Goldberg R, Pham QL, Giang LM. Handling missing data in COVID-19 incidence estimation: Secondary data analysis. JMIR Public Health and Surveillance. 2024;10:e53719.
[13] Rendana M, Idris WMR, Rahim SA. Effect of COVID-19 movement control order policy on water quality changes in Sungai Langat, Selangor, Malaysia within distinct land use areas. Sains Malaysiana. 2022;51(5):1587–98.
[14] Zainuri NA, Jemain AA, Muda N. A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana. 2015;44(3):449–56.
[15] Sukatis FF, Noor NM, Zakaria NA, Ul-Saufie AZ, Suwardi A. Estimation of missing values in air pollution dataset by using various imputation methods. International Journal of Conservation Science. 2019;10(4):791–804.
[16] Libasin Z, Ul-Saufie AZ, Ahmat H, Shaziayani WN. Single and multiple imputation method to replace missing values in air pollution datasets: A review. IOP Conference Series: Earth and Environmental Science. 2020;616(1):012021.
[17] Libasin Z, Fauzi WSWM, Ul-Saufie AZ, Idris NA, Mazeni NA. Evaluation of single missing value imputation techniques for incomplete air particulates matter (PM10) data in Malaysia. Pertanika Journal of Science and Technology. 2021;29(4):3099–112.
[18] Chen M, Zhu H, Chen Y, Wang Y. A novel missing data imputation approach for time series air quality data based on logistic regression. Atmosphere. 2022;13(7):1044.
[19] Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesthesia and Analgesia. 2018;126(5):1763–8.