DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 5, ISSUE 2, 2025, 17 – 22
DOI: https://doi.org/10.15282/daam.v5i2.9725

RESEARCH ARTICLE

The imputation of missing tides data in Malaysian tourism areas using basic statistical methods

N.Z.A. Hamid*, A. Sahrin, 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 analyses the imputation of missing tide data in Malaysian tourism areas using basic statistical methods. It aims to determine the most appropriate method among five basic statistical methods for the imputation of missing tide data in three Malaysian tourist areas, namely Kota Kinabalu, Penang and Langkawi Island. These methods are Top Bottom Mean, 6-Hour Mean, 12-Hour Mean, Daily Mean, and Linear Interpolation. The data were recorded hourly for 14 days, which is equivalent to 336 hours, in 2019. The data are complete and continuous. The percentage of data discarded in this study is 10%, 20%, 30%, 40%, and 50%. The performance indices used to evaluate the methods are Correlation Coefficient (CC), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Overall, the best basic statistical method to impute missing tide data is Linear Interpolation, and it is hoped that this study can help the Department of Survey and Mapping Malaysia (JUPEM) in imputing the missing tide data.

ARTICLE HISTORY
Received : 9th Sept 2023
Revised : 28th Feb 2024
Accepted : 30th April 2024
Published : 30th Sep 2024

KEYWORDS
Imputation method
Missing data
Basic statistical method
Tides data
Malaysian tourism area

1. INTRODUCTION

Missing values in data can pose significant challenges, particularly in time series analysis, where a continuous and complete dataset, one without gaps or interruptions in recorded values, is essential for accurate predictions [1–4]. Therefore, this study focuses on the imputation of missing data. The missing data imputation methods employed were basic statistical methods, such as Linear Interpolation, Top Bottom Mean, Daily Mean, 12-Hour Mean, and 6-Hour Mean. To compare the performance of the methods, the performance indices used are Correlation Coefficient (CC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

Marine parks in Sabah are famous for their islands, sandy beaches, coral reefs, and clear waters. Sabah has at least seven marine parks gazetted by the state government specifically to promote the tourism industry [5]. According to [6], in Malaysia, coastal areas receive an exceptionally large number of tourists. Next, one of the states in Malaysia that is well-known for its coastal areas is Penang. Meanwhile, Langkawi Island varies with its natural beauty and unique history, which makes it a tourist destination that often attracts local and foreign tourists [7].

Therefore, it is essential to manage tidal patterns in popular tourist destinations such as Kota Kinabalu, Penang, and Langkawi Island to ensure the safety and sustainability of tourism activities. Tides influence various coastal operations, including beach accessibility, recreational water activities, and the maintenance of local infrastructure. However, tidal data may contain missing values due to factors such as equipment failure, extreme weather events, or data transmission issues. These gaps can hinder accurate tidal forecasting, potentially posing risks to tourists and affecting tourism-related planning. Thus, this study aims to address these challenges by improving the accuracy and reliability of tidal data, which is crucial for informed decision-making and the sustainable development of coastal tourism areas.

2. METHODOLOGY

2.1 Tides data

The time series used in this study is the tide data observed hourly from 18th March to 31st March 2019. According to [8–9], the characteristics of Malaysia's climate include a uniform temperature, high humidity and heavy rain. Malaysia has a cloudless sky even during the severe drought. Malaysia also rarely has a period of several days with no sunlight at all, except during the northeast monsoon season. Although the wind in Malaysia is generally weak, there are periodic changes in wind patterns. Based on these changes, four seasons can be distinguished: southwest monsoon, northeast monsoon and two shorter transitional monsoon seasons.

The southwest monsoon usually begins in the latter half of May or early June and ends in late September. The prevailing wind is generally from the southwest with a weak speed that is below 15 knots. The northeast monsoon usually starts in early November and ends in March. During this season, the prevailing wind is from the east or northeast with a speed between 10 and 20 knots. The northeast monsoon is very significant, with heavy rains and rough tides. For this reason, this study chose the period of the northeast monsoon, which is the month of March.

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

The selected study locations are the Malaysian tourist spots of Kota Kinabalu, Penang and Langkawi Island. These three areas are the focus of tourists, and the tide level needs to be focused on to predict any disaster that may occur. The tides data are provided by the Department of Survey and Mapping Malaysia (JUPEM) [10]. These tide data are used for various surveying and mapping activities, development, planning, navigation, disaster management and scientific studies. The entire duration of the time series is 336 hours, recorded in units of centimetres (cm) above the Tide Gauge Zero.

2.2 Imputation of missing data

This study imputes missing tide data from 2019 by simulating missing data at rates of 10%, 20%, 30%, 40%, and 50%. The percentages up to 30% are typically examined in existing literature, while 40% and 50% are included to further assess the robustness and efficiency of the imputation method under higher levels of missingness, which are typically less explored but essential for evaluating method performance in more extreme cases. Through the percentage of missing data, data calculations using five basic statistical methods were carried out. The missing data imputation methods were compared and measured based on performance index calculations to determine the best of the five methods.

2.2.1 Basic statistical methods

Basic statistical methods were applied to impute the data that had been randomly discarded by 10%, 20%, 30%, 40%, and 50%. The basic statistical methods are Linear Interpolation, Top Bottom Mean, Daily Mean, 12-Hour Mean, and 6-Hour Mean. Most of the imputation methods have been used by previous research, such as [1, 11–15]. However, most of the research applies the methods to estimate the missing values of the air pollution dataset. This present study attempts to apply the methods to tide data.

Linear Interpolation (LI): Linear interpolation connects two data points with a straight line. Therefore, the missing value can be calculated directly using a linear equation. The equation is written as:

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

where y^* is the missing data, x^* is the time point of the missing data, x_1 and y_1 are the coordinate of the starting point of the missing data, and x_2 and y_2 are the coordinate of the final point of the missing data.

Top Bottom Mean (TBM): This method calculates the missing data using the mean calculation of the point above and the point below the missing data. The equation is as follows:

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

Daily Mean (DM): The daily mean is the mean of the observation data calculated every 24 hours from hour 1 to 336. The missing data will be imputed with the 24-hour mean data. The equation is represented as below:

$$\bar{y} = \frac{\sum_{j=1}^{24} y_j}{24} \tag{3}$$

where $\bar{y}$ is the missing data and y_j is the data of the tide height at the j^{th} hour. For example, if the missing data is at the 50th hour, then $\bar{y}$ will be imputed with the average of the complete tidal height data from the 12 hours before and after the discarded data.

12-Hour Mean (M12): The missing data were calculated with the 12-hour average values. The equation is as follows:

$$\bar{y} = \frac{\sum_{j=1}^{12} y_j}{12} \tag{4}$$

6-Hour Mean (M6): Similar to the method in the previous section, the 6-hour average is the average of observation data calculated every 6 hours. The equation is:

$$\bar{y} = \frac{\sum_{j=1}^{6} y_j}{6} \tag{5}$$

2.3 Performance index

In order to find out the best method for the five basic statistical methods, the performance index was used in this study. The missing data that had been imputed and the original data for the missing data were compared to determine the best method for replacing the missing data. The three performance index methods used in this study are CC, MAE, and RMSE.

Correlation Coefficient (CC): This indicator describes the variability in the calculated data and how much it relates to the observed data. It takes a value between 0 and 1, with a value closer to 1 implying a better fit. The equation can be expressed as:

$$CC = \frac{\sum_{i=1}^{N}(P_i - \overline{P})\,(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2}\ .\ \sqrt{\sum_{i-1}^{N}(O_i - \overline{O})^2}} \tag{6}$$

where N is the total of the imputed data, P_i is the imputed data, O_i is the original data, $\overline{P}$ is the mean of the imputed data, and $\overline{O}$ is the mean of the original data.

Mean Absolute Error (MAE): The MAE is the average difference between the calculated and observed data and is given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{7}$$

Root Mean Square Error (RMSE): RMSE is one of the most common methods for evaluating numerical predictions. The value is calculated with the equation:

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}} \tag{8}$$

A smaller RMSE value indicates a better performance of the method.

3. STUDY FINDINGS AND DISCUSSION

3.1 Forecasting results

The total interval of time series data for tides in Kota Kinabalu, Penang and Langkawi Island is 336 hours. The time series data that had been discarded randomly by 10%, 20%, 30%, 40%, and 50% were calculated using the five basic statistical methods and then tested using the performance index to determine the best method. Table 1a-1e shows the performance index results of five basic statistical methods for 10%, 20%, 30%, 40%, and 50% of the tidal missing data in Kota Kinabalu, respectively. Table 2a-2e displays the performance index results of five basic statistical methods for 10%, 20%, 30%, 40%, and 50% of the missing tide data in Penang. Next, Table 3a-3e demonstrates the performance index results of five basic statistical methods of 10%, 20%, 30%, 40%, and 50% of the tidal missing data on Langkawi Island, respectively.

Table 1a. Performance index results of five basic statistical methods for 10% missing data at Kota Kinabalu

Method	CC	MAE	RMSE
LI	0.99	3.49	4.45
TBM	**0.99**	**3.31**	**4.28**
DM	0.03	33.30	38.18
M12	0.63	25.70	30.57
M6	0.96	12.13	15.09

Table 1b. Performance index results of five basic statistical methods for 20% missing data at Kota Kinabalu

Method	CC	MAE	RMSE
LI	**0.99**	**3.99**	**5.13**
TBM	0.99	4.65	6.05
DM	0.02	30.28	36.05
M12	0.46	26.9	31.58
M6	0.90	14.33	16.87

Table 1c. Performance index results of five basic statistical methods for 30% missing data at Kota Kinabalu

Method	CC	MAE	RMSE
LI	**0.99**	**4.62**	**5.61**
TBM	0.98	6.29	8.64
DM	0.22	33.44	39.62
M12	0.56	26.09	31.37
M6	0.90	15.23	18.5

Table 1d. Performance index results of five basic statistical methods for 40% missing data at Kota Kinabalu

Method	CC	MAE	RMSE
LI	0.91	**7.75**	16.25
TBM	**0.93**	8.40	**13.01**
DM	0.09	32.48	38.15
M12	0.43	27.21	32.20
M6	0.81	18.08	21.57

Table 1e. Performance index results of five basic statistical methods for 50% missing data at Kota Kinabalu

Method	CC	MAE	RMSE
LI	**0.98**	**5.95**	**7.92**
TBM	0.88	11.00	16.30
DM	0.10	29.94	36.08
M12	0.44	26.54	31.69
M6	0.81	17.66	21.16

Table 2a. Performance index results of five basic statistical methods for 10% missing data at Penang

Method	CC	MAE	RMSE
LI	**0.99**	**10.06**	**14.36**
TBM	0.98	11.92	18.48
DM	0.11	51.12	61.73
M12	0.67	55.91	69.17
M6	0.93	30.62	39.69

Table 2b. Performance index results of five basic statistical methods for 20% missing data at Penang

Method	CC	MAE	RMSE
LI	**0.99**	**9.68**	**13.22**
TBM	0.98	11.36	15.87
DM	0.00	53.88	64.48
M12	0.60	62.14	76.15
M6	0.79	36.86	46.41

Table 2c. Performance index results of five basic statistical methods for 30% missing data at Penang

Method	CC	MAE	RMSE
LI	**0.98**	**9.58**	**13.82**
TBM	0.91	14.81	24.71
DM	0.17	46.37	57.70
M12	0.45	51.34	65.20
M6	0.74	33.42	42.98

Table 2d. Performance index results of five basic statistical methods for 40% missing data at Penang

Method	CC	MAE	RMSE
LI	**0.97**	**13.54**	**19.37**
TBM	0.93	16.03	23.63
DM	0.07	49.00	61.23
M12	0.31	51.48	65.42
M6	0.51	40.13	52.43

Table 2e. Performance index results of five basic statistical methods for 50% missing data at Penang

Method	CC	MAE	RMSE
LI	**0.79**	**22.12**	**35.93**
TBM	0.73	25.38	39.51
DM	0.02	46.24	57.44
M12	0.39	50.96	65.48
M6	0.26	41.97	56.52

Table 3a. Performance index results of five basic statistical methods for 10% missing data at Langkawi Island

Method	CC	MAE	RMSE
LI	**0.99**	**10.01**	**12.01**
TBM	0.99	11.15	13.90
DM	0.04	68.50	81.76
M12	0.64	78.34	93.48
M6	0.95	39.26	48.31

Table 3b. Performance index results of five basic statistical methods for 20% missing data at Langkawi Island

Method	CC	MAE	RMSE
LI	0.89	14.64	32.94
TBM	**0.98**	**12.04**	**17.37**
DM	0.44	54.43	67.38
M12	0.58	67.66	86.02
M6	0.78	38.53	50.85

Table 3c. Performance index results of five basic statistical methods for 30% missing data at Langkawi Island

Method	CC	MAE	RMSE
LI	**0.97**	**15.43**	**24.05**
TBM	0.95	16.91	26.16
DM	0.01	61.23	75.24
M12	0.64	70.35	88.14
M6	0.70	45.05	58.15

Table 3d. Performance index results of five basic statistical methods for 40% missing data at Langkawi Island

Method	CC	MAE	RMSE
LI	**0.91**	**21.58**	**33.24**
TBM	0.79	28.62	44.57
DM	0.17	62.33	78.40
M12	0.42	66.75	84.68
M6	0.27	54.23	71.27

Table 3e. Performance index results of five basic statistical methods for 50% missing data at Langkawi Island

Method	CC	MAE	RMSE
LI	**0.84**	**27.85**	**43.48**
TBM	0.80	29.68	46.61
DM	0.06	59.99	76.30
M12	0.36	65.45	83.42
M6	0.22	56.92	75.67

4. CONCLUSIONS

Basic statistical methods can be used to impute the missing tide data in three Malaysian tourism areas. It can be concluded through this study that the best method among the five basic statistical methods for missing tide data imputation is LI. Therefore, this missing data imputation method can be proposed to predict the tide time series data.

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. An utmost appreciation to the Department of Survey and Mapping Malaysia (JUPEM) for sharing the data.

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

Individual Assistant
NA

AUTHOR CONTRIBUTIONS

N.Z.A. Hamid (Conceptualization; Formal analysis; Writing- original draft), A. Sahrin (Conceptualization), N.B.A. Wahid (Methodology); N.H. Adenan (Formal analysis), N.H.M. Husin (Methodology), N.W.M. Junus (Formal analysis), 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] 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.
[2] Zakaria NA, Noor NM. Imputation methods for filling missing data in urban air pollution data for Malaysia. Urbanism. 2016;9(2):159–66.
[3] Noor NM, Yahaya AS, Ramli NA, Al Bakri Abdullah MM. Filling the missing data of air pollutant concentration using single imputation methods. Applied Mechanics and Materials. 2015;754–755:923–32.
[4] Saeipourdizaj P, Sarbakhsh P, Gholampour A. Application of imputation methods for missing values of PM10 and O3 data: interpolation, moving average, and k-nearest neighbor methods. Environmental Health Engineering and Management. 2021;8(3):215–26.
[5] Johnes J, Mapjabil J. Pola ruang keliaran pelancongan kembara di Kota Kinabalu, Sabah. Journal of Tourism, Hospitality & Environmental Management. 2020;5:242–54.
[6] Azhar N, Ahmad H. Beach tourism and family tourists satisfaction: A case of Penang. Jurnal Wacana Sarjana. 2019;3(1):1–14.
[7] Samad S, Shukor MS, Salleh NHM. Impak pembangunan industri perlancongan kepada komuniti di Pulau Langkawi. Proceedings of the Eighth National Economics Conference of Malaysia (PERKEM VIII). 2013;1:207–16.
[8] Malaysian Meteorological Department. Iklim Malaysia. 2023.
[9] Malaysian Meteorological Department. Fenomena Cuaca. 2023.
[10] Department of Survey and Mapping Malaysia. Garis Panduan Teknikal Cerapan Air Pasang Surut [Internet]. 2021 [cited 2025 May 9]. Available from: https://www.jupem.gov.my/storage/upload/pekeliling/23541-pkpup-8-2021.pdf
[11] 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.
[12] Ghapor AA, Zubairi YZ, Imon AHM. Missing value estimation methods for data in linear functional relationship model. Sains Malaysiana. 2017;46(2):317-26.
[13] 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) 012002.
[14] 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 & Technology. 2021;29(4):3099–3112.
[15] 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 (Basel). 2022;13(7)1044.
[16] Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia. 2018;126(5):1763–8.