Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 8, 2022, page 209-216

Missing Values Imputation for Monthly Wind Speed Data at Senai Johor

Siti Rokaiyah Sahidon*, Nur Arina Bazilah Kamisan
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
*Corresponding author: rokaiyah@graduate.utm.my

Abstract
Missing values are one of the most common issues that arise throughout the data observation or data recording process. Missing data is prevalent and can have a substantial impact on the inferences that can be formed from the data. It can cause various problems. Therefore, this research aimed to solve this problem using several simple methods in order to replace the missing values in the data without affecting the percentage of missingness. In this research, the data used is monthly data of maximum wind speed data from Senai, Johor that recorded from January 2009 until December 2019. The data is obtained from Malaysia Meteorological Department (MetMalaysia). In order to find the solution, three simple methods are being proposed; mean imputation, multiple imputation and linear regression to replace the missing values in the dataset. by using root-mean-square error (RMSE) and mean absolute percentage error (MAPE), the percentage of the missingness for each method are being compared. Linear regression is the most suitable method to use in order to impute the missing values for this data based on the lowest value for statistical error.

Keywords Missing values; wind speed; mean imputation; multiple imputation; linear regression

1. Introduction

The study of missing data began in 1960, and it has grown dramatically across numerous subject areas since then (Adibah, 2020). Missing values are one of the most common issues that arise throughout the data observation or data recording process (Pratama, 2016). Missing data, also known as missing values in statistics, arise when no data value for a variable in an observation is recorded.

Missing data is prevalent and can have a substantial impact on the inferences that can be formed from the data. Nonresponse can result in missing data: no information is supplied for one or more components or for the entire unit ("subject"). Items regarding private matters, such as wealth, are more likely to elicit a nonresponse than others. Attrition is a sort of missingness that can occur in longitudinal research, such as developmental studies, when a measurement is repeated after a given amount of time. Missingness happens when a person drops out before the finish of the test, leaving one or more measures unrecorded (“Missing data”, 2020).

In this, data that we use is maximum wind speed data that contained missing values and needed to solve. Wind speed data being used in order to solve the problems, some methods can be used but it will need to be test if the type of missingness mechanism will be suitable with the method that will be use.

2. Literature Review

2.1 Introduction

Missing values are one of the most common issues that arise throughout the data observation or data recording process the process of data observation or data recording. The demand of the data completion through the data observation for the advanced analysis becomes critical to be solved. The issue of missing data is inherent while performing data gathering studies. The study of missing data began in 1960, and it has grown dramatically across numerous subject areas since then (Adibah, 2020).

Typically, data is unavailable or missing owing to a variety of factors such as incorrect data entry, network errors, equipment malfunction, and database system issues. This problem needs to be deal wisely and solved carefully.

Missing values can be categorized into three parts (Pratama, 2016). The first part is missing completely at random (MCAR). For this kind, the variable is absent fully at random, with no missingness likelihood requirements tied to the variable itself. The second portion is missing at random (MAR), which means that the variable is missing at random if the likelihood of missingness is solely determined by available information. The third component is not missing at random (NMAR), which explains how the missingness probability varies based on the variable.

Missing values are classified into two types based on their absence. The first pattern is that the data are monotone absent. If we see a pattern in the missing data, we may need to restructure or rearrange variables or people. The second pattern is missing arbitrarily if the data may be reordered or rearranged to produce the obvious pattern of missingness.

2.2 History of missing data

The author state that the research on missing data was initiated in 1960 and it grew exponentially across various subject areas until now. From this paper, they aim to analyze the context of missing data. In order to conduct research that related to data collecting, missing data is unavoidable. The author state that the chances of observational research to encounter this problem is almost certain and extra care needed in handling it

Several times during the past 60 years, the publishing of papers regarding missing data has grown dramatically, but the most dramatic increase occurred in 2016, with 446 publications compared to 361 in 2015. Based on the observation that, as of December 2019, publications indexed for 2020 already total 24 articles, it is expected that further publications in the context of missing data would be released in 2020 (Farah Adibah et. al, 2020).

Furthermore, this study focuses solely on the issue of missing data, based on the title and author keyword used in the articles. Furthermore, the search was limited to journal articles alone, ruling out all other documents and source categories related to missing data.

2.3 A review of missing values handling methods on time-series data

Several methods have been introduced to solve missing values according to its missing mechanism. Conventional methods such as mode and mean imputation and deletion are approaches that arose early in the missing values handling studies and are generally basic yet dangerous methods, believed to be risky since they solve one problem while introducing another (Pratama et. al, 2016). The second method is imputation procedures. Imputation, often known as estimation, is a process for dealing with missing data by replacing each missing value with a specific value. The values source would be different for each method. It is stated that the percentage of missingness will affect the method we use to solve every missing values problem (Wu et. al, 2015 and Sridevi et. al, 2011).

2.4 Handling missing values in longitudinal panel data with multiple imputation

In the field of family research, Young et. al (2015) stated that longitudinal panel (prospective) survey data is commonly used. From 2010 to 2014, the Journal of Marriage and Family published roughly 287 quantitative and qualitative research publications (excluding theory development, research reviews, comments, rejoinders, and methodological innovation pieces) (JMF).

In this paper, the author looked at how missing data might be handled in two regression models, fixed effects and event history, because they are typically produced by researchers using a data structure where repeated observations are nested within individual records. They examined whether it is possible to impute both within-wave and whole-wave missingness.

It is turned out imputation improved estimates in the event-history analysis but only modest improvements in the estimates and standard errors of the fixed effects analysis. After that, they examine the factors that responsible for the differences in the value of imputation.

3. Methodology

3.1 Description of the data

The monthly data of maximum wind speed data from Senai, Johor recorded from January 2009 until December 2019. The data obtained is from Malaysia Meteorological Department (MetMalaysia).

3.2 Method of missing value imputation

3.2.1 Multiple imputation

Multiple imputation is a broad solution to the problem of missing data that is accessible in a number of widely used statistical tools. It seeks to account for uncertainty about missing data by generating numerous plausible imputed data sets and suitably integrating the findings gained from each of them.

This method uses to fill in the missing values multiple times, creating multiple "complete" datasets. The values those are missing are imputed based on the observed values for a specific individual and data relationships seen for other participants, providing that the observed variables are included in the imputation model.

In missingness mechanisms, multiple imputation can be used when data is missing totally at random, missing at random, and missing not at random. One of the benefits of employing multiple imputation is that it is versatile and may be applied to a wide range of settings. It is demonstrated that we may use this strategy to identify missing values in wind speed data.

$$\boldsymbol{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i} \tag{1}$$

$$\bar{x} = Average\ of\ the\ data$$
$$N = Total\ of\ data$$
$$x_i = Value\ of\ the\ data$$

3.2.2 Mean imputation

Mean imputation method is one of the simple methods to find the missing data in a dataset. The mean of the observed values for each data being compute and the missing data in the dataset are imputed by this value. This method maintains the sample size and easy to use, but the variability in the data is reduced, so it will affect the standard deviation and the variance estimates tend to be underestimated.

It is said that by using this method towards Missing Completely at Random (MCAR) type of data, it will lead to a severe biased estimate. It is clearly show that if the number of the missing data in a variable is large and these values are imputed by the sample mean, the resulting variance estimate will be underestimated.

This method being used by impute the average mean for the whole datasets and filling it into the missing part for each data. Another option that can be use is by impute the average of each data which is by calculating it by months, years or any classes of data in the datasets. Both of these methods will result biased analysis especially when the missing data are not MCAR.

$$\boldsymbol{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i} \tag{2}$$

$$\bar{x} = Average\ of\ the\ data$$
$$N = Total\ of\ data$$

$$x_i = Value\ of\ the\ data$$

3.2.3 Linear regression

Linear regression is a method that being used to predict the value of a variable based on the value of another variable. It is an approach for modelling the relationship between a scalar response and one or more explanatory variables.

For this method, the missing values can be filled by using Excel formula. The formula that being used is

$$\boldsymbol{Y_i = f(X_i, \beta) + e_i} \quad (3)$$
$$Y_i = Dependent\ variable$$
$$f = Function$$
$$X_i = Independent\ variable$$
$$\beta = Unknown\ parameters$$
$$e_i = Error\ terms$$

Linear regression usually used for determining the strength of predictors. It can be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Secondly, forecasting an effect or impact of changes. The regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. The last one is used for predicts trends and future values.

This method creates a simple model (a line) where it is easy to extrapolate or interpolate the missing values. It is only suited for the data that is likely to be linear. It is said that linear regression assumes the relationship between the independent and dependent variable is linear, so the line of best fit through the data points is a straight line rather than a curve.

Linear regression often can be used to predict the value of the dependent variable at certain values of the independent variable but it cannot predict the values that it outside the range of the data that being measured.

3.3 Performance indicator

3.3.1 Mean absolute error (MAE)

$$\boldsymbol{MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}} \quad (4)$$
$$y_i = observed\ value$$
$$x_i = predicted\ value$$
$$n = total\ number\ of\ observations$$
$$\Sigma = greek\ symbol\ means\ "sum"$$

3.3.2 Mean absolute percentage error (MAPE)

$$\boldsymbol{MAPE = \frac{1}{n}(\sum_{i=1}^{n} \left|\frac{y_i - x_i}{y_i}\right| \times 100} \quad (5)$$
$$y_i = observed\ value$$
$$x_i = predicted\ value$$
$$n = total\ number\ of\ observations$$
$$\Sigma = greek\ symbol\ means\ "sum"$$

3.3.3 Root mean square error (RMSE)

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \qquad (6)$$

$$y_i = observed\ value$$
$$x_i = predicted\ value$$
$$n = total\ number\ of\ observations$$
$$\Sigma = greek\ symbol\ means\ "sum"$$

4. Results and discussion

Three different types of methods with three different percentage of missing values towards the monthly wind speed data being used to compare which one is the best method that can be consider in finding missing values of the datasets. In order to find the best fit method towards the data, it is divided by three percentage of missingness which is 10%, 20% and 30% of missingness.

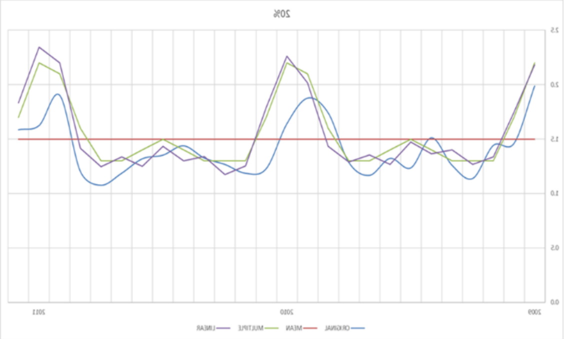

Figure 1 20% of missingness

The analysis of the results obtained from the values each of the methods impute and made in two directions, qualitative and quantitative. It compares the results from the original data, mean imputation method, multiple imputation method, and linear regression method.

From the graphs, we can conclude that 20% of missingness for multiple imputation is the most compatible among the three methods.

METHOD	MEAN IMPUTATION			MULTIPLE IMPUTATION			LINEAR REGRESSION		
PERCENTAGE	10%	20%	30%	10%	20%	30%	10%	20%	30%
MAPE=	17.7197	16.6955	17.4167	16.9979	**12.6623**	9.5008	**15.0919**	13.9716	**9.2331**
RMSE=	**0.2542**	0.4169	0.381	0.3113	**0.2449**	**0.1925**	0.292	0.2735	0.2176
MAE=	0.2308	0.2269	0.2821	0.2308	**0.1769**	0.1556	**0.2125**	0.1992	**0.154**

Table 1 Value error measures

Table 4.1 summarizes the performance results obtained from the proposed methods. There are two statistical indicators that being used to measure the performance of each method. These two indicators investigate the degree of precision. The lower the value of the indicator, the more accurate the predictive effect of the method

From Table 4.1, it is shows that the linear regression and multiple imputation compute the smaller error than mean imputation. From this table, we can say that there is not particular method that can be used to find the missing values. Instead, we can see that multiple imputation is show that it can be used to find the missing values in 20% missingness of the data.

Linear regression also shown as the most suitable method in finding the missing values for smallest percentage of missingness than the other two methods in the value of error measures.

Based on Table 4.1, multiple imputation is shown as the most suitable method to use in finding the missing values towards the 20% missingness of the data as the value of all error indicator for this method is the smallest among three of the methods.

Conclusion

Missing values are a common issue in the actual database. Thus, many methods have been used to deal with missing values. the most important thing is one should choose the appropriate method by determine the type of missing values that occur. It is because by using a suitable method, the missing values that being impute will be closest to the actual values.

The data that being used is a seasonal data that it is can be shown that it has a trend in the data. Then, this data is predictable that by imputing the missing values in here might be challenging due to the seasonality of the data.

By reduce the seasonality of the data into months, the complexity of the data can be reduced and a simpler method could be applied. Three simple methods being proposed for this purpose which is mean imputation, multiple imputation and linear regression.

When considering imputation of missing values, the direction of the missing values in the data is equally essential. It can be considered as a trend factor when the data shows a trend. From this study, it can be concluded that linear regression provides the best estimation of the missing values if the missing values are 10% and 20% and multiple imputation gives the best estimation for 20% missingness of the data. It is shows that the percentage of missingness should be taken into consideration before imputing the missing values.

To support the analysis results, the data being reduce into monthly data and the graph that we obtained show that the data is nearly to be linear. In conclusion, it is important to understand the type of data that being used before applying any methods to find the missing values in the data. By comprehending the data, the data's complexity, such as seasonality and trend effects, may be separated, and a simpler strategy for dealing with missing numbers can be employed.

Acknowledgement

In preparing this thesis, I was in contact with many people besides my own supervisor. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my supervisor, Dr Nur Arina Bazilah Binti Kamisan for the encouragement, guidance, critics and advice. Without her support and interest to help me finish my thesis, this would not have been the same as presented here.

I am also indebted with my fellow friends that always be supportive and offer helps whenever I need. My sincere appreciation also extends to all my fellow undergraduate student that should be recognized for their endless support. Their views and tips are very useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

References

[1] Álvarez, J., Allen, H. L., Albaugh, T. J., Stape, J. L., Bullock, B. P., & Song, C. (2013). Factors influencing the growth of radiata pine plantations in Chile. Forestry, 86(1), 13-26.

[2] A. S. Yahaya, N. A. Ramli, F. Ahmad, N. M. Nor, M. N. H. Bahrim, “Determination of the Best Imputation Technique for Estimating Missing Values when Fitting the Weibull Distribution”, International Journal of Applied Science and Technology, Vol. 1, no. 6, pp. 278 – 285, 2011

[3] A. T. Sree Dhevi, “Imputing missing values using Inverse Distance Weighted Interpolation for time series data,” 6th Int. Conf. Adv. Comput. ICoAC 2014, pp. 255–259, 2015.

[4] Adnan, Farah & Zakaria, Mohd & Ibrahim, Safwati. (2020). 60-year Research History of Missing Data: A Bibliometric Review on Scopus Database (1960-2019)

[5] B. Ran, H. Tan, Y. Wu, and P. J. Jin, “Tensor based missing traffic data completion with spatial–temporal correlation,” Phys. A, Statist. Mech.

[6] Barrios, A., Trincado, G. & Garreaud, R. Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. For. Ecosyst. 5, 28 (2018). https://doi.org/10.1186/s40663-018-0147 Appl., vol. 446, pp. 54–63, Mar. 2016.

[7] Cannell, M. G. R., Cruz, R. V. O., Galinski, W., Cramer, W. P., Alvarez, A., Austin, M. P., ... & Xu, D. (1996). Climate change impacts on forests. In CLIMATE CHANGE 1995 Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses (pp. 95-129).

[8] Central Agency on Statistics of Makassar city. Makassar in Figure 2010. Makassar: UD Areso, 2010.

[9] D. Little, RJA and Rublin, "Statistical Analysis with Missing Data," Wiley, New York., p. 381, 1987.

[10] De Souto, M. C. P., Jaskowiak, P. A., Costa, I. G., Bioinformatics 16 (2015) 64–72.

[11] Deng (2012-10-05). "On Biostatistics and Clinical Trials". Archived from the original on 15 March 2016. Retrieved 13 May 2016

Gerding, V., & Schlatter, J. E. (1995). Variables y factores del sitio de importancia para la productividad de Pinus radiata D. Don en Chile. Bosque, 16(2), 39-56.

[12] Ghorbani, S., Desmarais, M. C., Appl Artif Intell 31, 1 (2017) 1–22.

[13] Graham J. W. (2009). Missing data analysis: making it work in the real world. Annual review of psychology, 60, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

[14] H. Tan et al., "A tensor-based method for missing traffic data completion," Transp. Res. C, Emerg. Technol., vol. 28, pp. 15–27, Mar. 2013.

[15] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, "Short-term traffic prediction based on dynamic tensor completion," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 8, pp. 2123–2133, Aug. 2016.

[16] Hyndman, Rob J.; Koehler, Anne B. (2006). "Another looks at measures of forecast accuracy". International Journal of Forecasting. 22 (4): 679–688. CiteSeerX 10.1.1.154.9771. doi: 10.1016/j.ijforecast.2006.03.001

[17] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Inf. Sci. (Ny)., vol. 233, pp. 25–35, 2013.

[18] J. Van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," Transp. Res. C, Emerg. Technol., vol. 13, nos. 5–6, pp. 347–369, Oct./Dec. 2005.

[19] Kalton, Graham (1986). "The treatment of missing survey data". Survey Methodology. 12: 1–16.

[20] Kamisan, N.A.B, Lee M.H, Hussin A.G & Zubairi Y.Z. 2020. Imputation Techniques for Incomplete Load Data Based on Seasonality and Orientation of the Missing Values, Sains Malaysiana 49(5) (2020): 1165-1174, doi: http://dx.doi.org/10.17576/jsm-2020-4905-22

[21] Kang, P., Neurocomputing 118 (2013) 65–78.

[22] L. Li, J. Zhang, Y. Wang and B. Ran, "Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 8, pp. 2933-2943, Aug. 2019, doi: 10.1109/TITS.2018.2869768.

[23] L. Li, S. He, J. Zhang, and B. Ran, "Short-term highway traffic flow prediction based on a hybrid strategy considering temporal–spatial information," J. Adv. Transp., vol. 50, no. 8, pp. 2029–2040, Dec. 2016.

[24] Mike W. & Jeff H. (2013). Bayesian Forecasting and Dynamic Models. Springer Series in Statistics. Springer.

[25] Molnar, Frank J.; Hutton, Brian; Fergusson, Dean (2008-10-07). "Does analysis using "last observation carried forward" introduce bias in dementia research?". Canadian Medical Association Journal. 179 (8): 751–753. doi:10.1503/cmaj.080820. ISSN 0820-3946. PMC 2553855. PMID 18838445.

[26] Oren Anava, Elad Hazan, and Assaf Zeevi. Online time series prediction with missing data. In Proceedings of the International Conference on Machine Learning, pages 2191–2199, 2015.

[27] Pampaka, M., Hutcheson, G., Williams, J. International Journal of Research & Method in Education 39, 1 (2016) 19-37.

[28] Pati, S. K., Das A. K., Knowl Inf Syst 52 3(2017) 709–750.

[29] Polit DF Beck CT (2012). Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed. Philadelphia, USA: Wolters Klower Health, Lippincott Williams & Wilkins

[30] Pontius, Robert; Thontteh, Olufunmilayo; Chen, Hao (2008). "Components of information for multiple resolution comparison between maps that share a real variable". Environmental Ecological Statistics. 15 (2): 111–142. doi:10.1007/s10651-007-0043-y

[31] Pratama, A. E. Permanasari, I. Ardiyanto and R. Indrayani, "A review of missing values handling methods on time-series data," 2016 International Conference on Information Technology Systems and Innovation (ICITSI), 2016, pp. 1-6, doi: 10.1109/ICITSI.2016.7858189.

[32] Rén, B., Pueyo, L., Chen, C.H., Choquet, É., Debes, J.H., Duchêne, G., Ménard, F., & Perrin, M.D. (2020). Using Data Imputation for Signal Separation in High-contrast Imaging. arXiv: Instrumentation and Methods for Astrophysics.

[33] Rudolph E. K. (1960). A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, 82(1):35–45.

[34] S. Sridevi, S. Rajaram, C. Parthiban, S. SibiArasan, and C. Swadhikar, “Imputation for the analysis of missing values and prediction of time series data,” 2011 Int. Conf. Recent Trends Inf. Technol., pp. 1158–1163, 2011.

[35] Sree Dhevi, A.T. (2014). Imputing missing values using Inverse Distance Weighted Interpolation for time series data. 2014 Sixth International Conference on Advanced Computing (ICoAC), 255-259.

[36] S. Wu, C. Chang, and S. Lee, “Time Series Forecasting with Missing Values,” 2015 1st Int. Conf. Ind. Networks Intell. Syst., pp. 151–156, 2015

[37] Swamidass, P. (2000) MEAN ABSOLUTE PERCENTAGE ERROR (MAPE). Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA. https://doi.org/10.1007/1-4020-0612-8_580

[38] Valdiviezo, H. C., Van, A. S., Inf Sci 31 1(2015) 163–181.

[39] Willmott, Cort; Matsuura, Kenji (2006). "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators". International Journal of Geographical Information Science. 20: 89–102. doi:10.1080/13658810500286976

[40] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. IEEE Computer, 42:30–37, 2009.

[41] Young, R., & Johnson, D. R. (2015). Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. Journal of marriage and the family, 77(1), 277–294. https://doi.org/10.1111/jomf.12144

[42] Yu H. F., Nikhil R. & Inderjit S. D. (2016). High-dimensional Time Series Prediction with Missing Value