Applied Mathematics and Computational Intelligence
Volume 10, No.1, Dec 2021 [319-327]

Missing Values Imputation For Wind Speed

Nur Arina Bazilah Kamisan[1]*, Siti Mariam Norrulashikin[2] and Siti Fatimah Hassan[3]

[1,2]Jabatan Sains Matematik, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor.
[3] Pusat Asasi Sains Universiti Malaya, Universiti Malaya, 51310 Kuala Lumpur.

* Corresponding author: nurarinabazilah@utm.my

Received: 25 October 2021; Accepted: 9 November 2021; Available online: 20 December 2021

ABSTRACT

Addressing missing values is important in the process of getting a precise and accurate result. If missing data are not treated appropriately, then the results could lead to biased estimates. But different series may require different strategies to estimate these missing values. Seasonal data has a repetitive cycle that is predictable. By disaggregating the data into it seasonal factors, clear information behavior of the data could be observed and will make it easier to deal with the missing value. In this paper, the performance of three different methods is being compared with each other. One of the imputation methods will used information from the seasonality for the missing values to enhance the imputation technique. the other two methods are mean interpolation and AR model as the missing values imputation. Wind speed data from Alor Setar, Malaysia are used for this purpose. From the error measurement, the enhanced technique gives the best performance compared to the other two techniques.

Keywords: autoregressive model, imputation, mean interpolation, missing values, wind speed.

1 INTRODUCTION

The need for power has increased as the world's population has expanded economically. A well-organized power operation system becomes more important to get a stable and sustainable electricity supply. Because the demand for power continues to rise, careful planning is required to improve the delivery of electricity to consumers [1]. One of the origins of energy is wind power. Wind power would harvest the flow of energy due to the uneven heating of the Earth's atmosphere by the Sun. Furthermore, wind energy is the easiest, most efficient, least polluting, and fastest-growing green energy source. Wind forecasting would be difficult due to diurnal, hourly, annual, and seasonal variations that vary drastically. A wind turbine is used to transform wind energy into electricity. As a result, wind speed data is being gathered to predict future energy generation [2].

In this full of technologies era, renewable energy resources such as solar energy and wind energy have been actively implemented in the last few years [3]. Wind power is one of the sources of electricity. To achieve the target of 20% of electricity generation from renewable energy sources, Malaysia will need RM33 billion investment to hit renewable energy target. There are various meteorological factors like the direction of air, air density, air pressure and wind speed can affect the wind power generator.

Additionally, factors like diurnal, hourly, annually and seasonal patterns change dramatically will make wind forecasting challenging [4]. To convert the wind energy into electricity, a wind turbine is used. For this reason, wind speed data are being collected to forecast the future production of electricity. Complete and high precision datasets need to be used to forecast the wind speed in order to obtain vital information from it.

As mentioned by Tawn, Browell and Dinwoodie [5], missing values in short term wind power forecasting always being neglected and lead to a bias estimation in generating forecasts. Other than that, missing data in wind power time series also could give impact on income, energy management and weakness estimation. Missing data or missing values pose a challenge for researchers who are trying to find a solution or assess missing wind speed values. Missing values in the data collection will influence the level of results on data classification, according to [6] so ignoring the missing values will lead to the incorrect inference.

Data that is collected after being performed under various conditions for many times and missing a few data due to certain issues may also be described as missingness process. Researchers all throughout the world recognise the importance of successfully dealing with missing data. Missing data, on the other hand, was frequently overlooked, and instances with partial missing data in variables included in the analysis were simply eliminated. This rule isn't appropriate for all domains [7]. According to De Goeij et al. [8] missingness mechanisms can be divided into three categories: missing totally at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

Before the analysis being done, it is essential to make sure the datasets are complete. There are many methods to resolve the missing data such as ignoring the missing data, deletion, and mean/mode imputation. One of the easiest methods is by simply ignore the missing data and only analyze the observed response but this results in loss of information and decreases the statistical power. Although this might be the simplest method, it is still enough to mislead the results of analysis because of the big missingness percentage Pratama et al. [9]. Although the simplest method to overcome the missing values problem is by using the mean imputation or unconditional mean imputation. This method completes the datasets by replacing the missing values with the mean of the observed data for the particular variable [9].

2 METHODOLOGY

In this study we will used four techniques to impute the missing values. The techniques are mean interpolation, linear interpolation, autoregressive (AR) model and an approach by using combination of mean interpolation with AR model. As mentioned by Choong, Charbit, & Yan [10], AR model is chosen because it has the advantage of being efficient for a time series data that contain missing values because it considers the dynamic nature of microarray temporal data as well as the data's local similarity structures.

2.1 Mean Interpolation

The most straightforward way to deal with missing values is to calculate the mean of the measured data depending on the precise timing. For example, to use the mean of the wind speed from 1st August 2013, 2014, 2015, and 2016 to impute the absent values of wind speed on 1st August 2017.

Strike, Emam, and Madhavji [11] also clarified the benefit of this approach is that it does not require any values or calculations.

2.2 Linear Interpolation

Moon et al. [12] (2019) explain that linear interpolation is a statistical method which is always being used to estimate with a function based on the known data. Linear interpolation is let the 2 known points with the coordinate (t_0, y_0) and (t_1, y_1). Linear interpolation is a straight line which will connect these 2 points. To obtain the missing value at y, equation of slope is playing an important role in it.

$$\frac{y - y_0}{t - t_0} = \frac{y_1 - y_0}{t_1 - t_0} \qquad (1)$$

where

y is the missing value at time t

y_0 is the wind speed at time t_0

y_1 is the wind speed at time t_1.

2.3 Autoregressive Model

Autoregressive model is a very random mechanism that is used to characterize time-varying processes [13]. Traditionally, according to Liu, Kumar and Palomar [14], AR was used to understand how time series operate, but it was also used to look for missing values due to computer malfunction or machine record failure. The model can be written as:

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t \qquad (2)$$

where

$\emptyset_1, \emptyset_2, \emptyset_p$ are the parameters of the model

ε_t is the white noise

p is the order of the autoregressive part.

According to Hassani and Yeganegi [15], the Liung-Box is used to pick the best p-values for the AR because it is very sensitive to the number of lags. This approach, however, can only determine if the p-values are consistent with the formula. As a result, Akaike's Information Criterion (AIC) is used to aid in the selection of the best model from many candidates. As a result, Akaike's Knowledge Criteria (AIC) is used to help in the selection of the best model from a range of p-values.

$$AIC = -2\log(L) + 2k \qquad (3)$$

where

L is the likelihood data

k is the number of parameter.

2.4 The combination of mean interpolation and AR model

The combination technique is approached to enhance the imputation. To accomplish this goal, the combination strategy is used to improve the precision of missing values imputation. First, we look for the missing values in an incomplete data collection. To fill in the missing values for temporary, we use the mean interpolation to substitute the missing values. Lastly, by using the AR model, we forecast or estimate the missing values in the data. The framework of this technique is shown in the Figure 1 below.

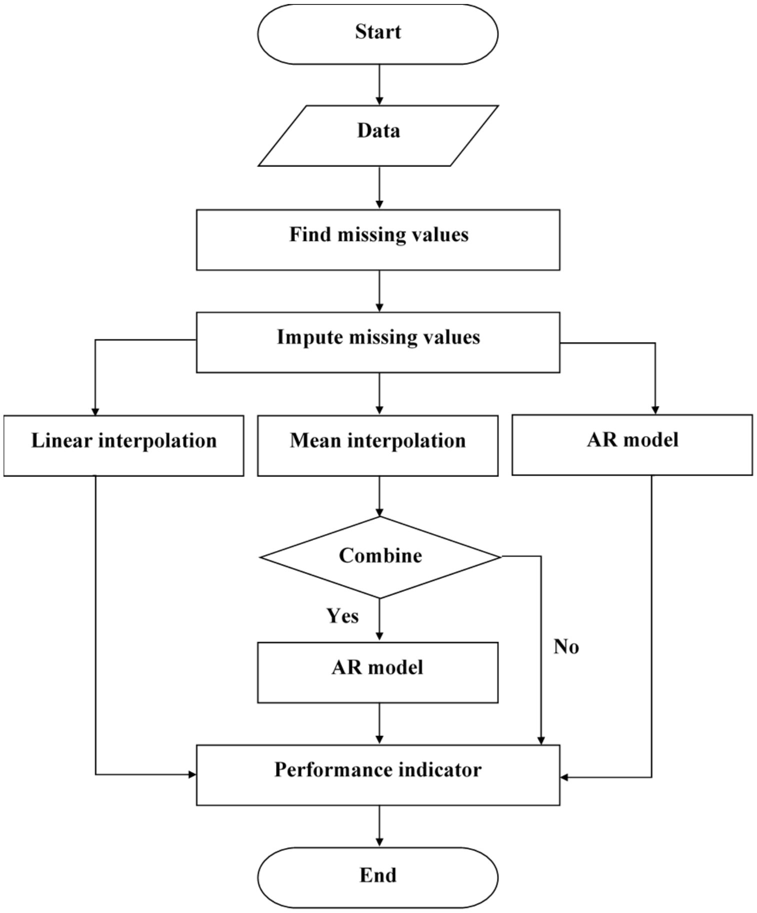

Figure 1 : Framework of the missing values imputation

The mean interpolation is a simple and efficient model for missing values. Thus, it is appropriate to apply this method as the temporary imputation for the missing values before being applied to AR model. The AR model is chosen because it can forecast the wind speed data well especially for in sample forecast.

2.5 Mean Absolute Percentage Error

The sum of the absolute error divided by its real value multiplied by 100 percent divided by n is known as MAPE. MAPE can be commonly used where the given data is sufficient to calculate precision, so it can represent the extent of missing values.

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{(y_t - \hat{y}_t)}{y_t}\right| \qquad (4)$$

where

y_t actual data at time t

$\hat{y}_t$ imputed missing values at time t

n number of missing data.

The following **Error! Reference source not found.** shows the interpretation of the MAPE values.

Table 1 : Interpretation of MAPE values

MAPE value (%)	Level of accuracy
MAPE ≤ 10	Highly Accurate
10 < MAPE ≤ 20	Good
20 < MAPE ≤ 50	Reasonable
MAPE ≥ 50	Inaccurate

3 RESULT AND DISCUSSION

Data from METMalaysia is daily maximum wind speed and direction data for Alor Setar. The time series plot of the Alor Setar can be seen from Figure 2 The data was compiled from January 1st, 2013 to December 31st, 2017. The missing wind speed data spans the months of August to November 2017. 10% missing values, 20% missing values, and 30% missing values are applied to the wind speed data for the purpose of this study.

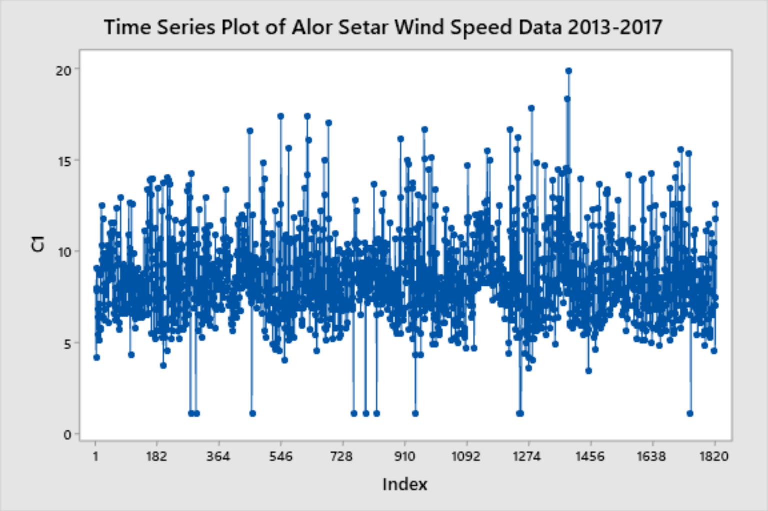

Figure 2 : Time series plot of Alor Setar

After the missing values being applied to the data, the mean interpolation and linear interpolation techniques are applied to impute the missing data. As for the AR model, the PACF plot is used to find the value of *p*. Since the PACF graph displays a decaying pattern in Figure 3 below, it means the data is stationary. Hence, the potential values for *p* are 1,2, and 4. Thus, three possible model are AR(1), AR(2) and AR(4). To check whether these models are adequate or not, we also used Ljung-Box test. If the *p*-value is larger than 0.05 it means the model is adequate. Lastly, to find the best model of the AR, we used the AIC. Smaller value of AIC indicates the model is better. In this study, AR(2) was chosen as an appropriate model because it has the smallest AIC value, which is 6.6108 less than AR(4).

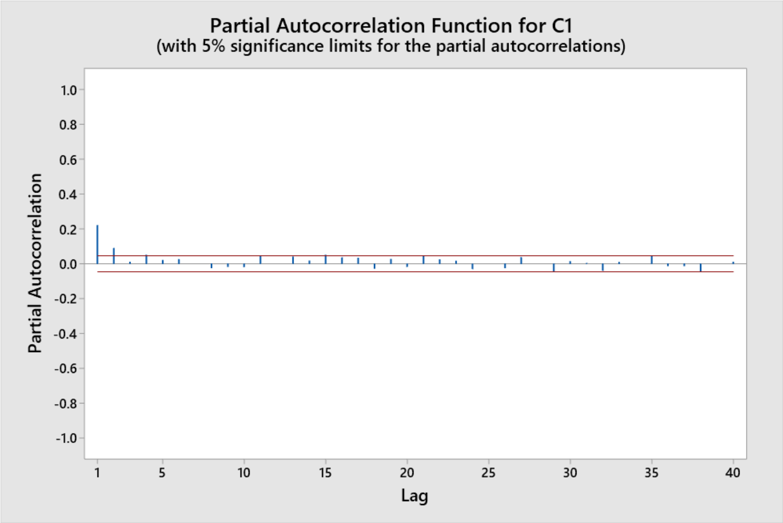

Figure 3 : PACF plot for Alor Setar that contain missing data

Table 2 : Model checking to find adequate AR model for the missing data

	***p*-value**	**Model Adequacy**	**AIC**
AR(1)	0.001	No	4.604
AR(2)	0.188	Yes	6.611
AR(4)	0.346	Yes	10.612

The same process is repeated for the missing data after being imputed with mean interpolation technique. The data are plotted with PACF plot as seen in figure below to find the *p*-value. After that Ljung-Box test is used to find adequate model and lastly AIC is used to find the best model of AR for each of selected missing percentage data. The result of the AIC comparison can be seen from Table 2 above. The results for the missing values can be seen in Table 3 below.

Table 3 : MAPE test to determine the best model

Technique	**MAPE**
Mean interpolation	2.556
Linear interpolation	3.301
AR model	2.323
Mean interpolation + AR model	2.286

From Table 3 above, the combination approach gives the smallest value of MAPE compared to the others. Although the difference is not very huge between the techniques used in this study, but it is reasonable because the data are recorded in small scale (m/s). Furthermore, a small improvement is still an improvement, and this technique has outperformed the other techniques used in this study.

By integrating the mean interpolation and AR model, it improves the precision of the imputed missing values while lowering the error. This technique first used the mean interpolation as the temporary substitution, and by referring to Table, it gives pretty small value of MAPE as missing value substitution. It later being analyze by using AR model which also gives a small MAPE in Table as missing value substitution. Therefore, by combining these technique and model will help improve the imputation of the missing values and makes it closer to the real data value.

4 CONCLUSION

Combination of two models or techniques have always shown that it could improve the outcomes. This is because each technique or model could always cover the limitations of the other method. In this case, AR help cover the mean interpolation limitation. Mean interpolation is a very simple approach with an adequate outcome, but it is not recommended since it can result in a large error in terms of covariance or correlation whilst the autoregressive model circumstances that the output variable is linearly dependent on its own previous values and on an imperfectly predictable term, resulting in a stochastic difference equation or recurrence relation. Therefore, by combining them together could help improve the imputation of the missing values. In conclusion, a combination of two methods or maybe more could help improve the imputation of the missing values. For future study, a combination of a more advantage model such as artificial neural network or fuzzy time series models with other missing values imputation such as maximum likelihood and multiple imputation could be used to improve the technique in substitute the missing values.

ACKNOWLEDGEMENT

This research is supported by the Universiti Teknologi Malaysia under grant QJ130000.2654.17J90.

REFERENCES

[1] S. B. Yaakob, S. H. M. Tahar and A. Ahmed, "Investment Planning Problem in Power System Using Artificial Neural Network," *Applied Mathematics and Computational Intelligence,* vol. 7, no. 1, pp. 13-22, 2018.

[2] E. Cadenas and W. Rivera, "Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN Model," *Renewable Energy*, vol. 35, pp. 2732-2738, 2010.

[3] K. Wannakam and S. Jiriwibhakorn, "Evaluation of Wind Energy Production using Weibull Distribution and Artificial Neural Networks," in *2018 International Conference on Engineering, Applied Sciences, and Technology*, 2018, pp. 1-4.

[4] Y. Zhang, S. J. Kim, and G. B. Giannakis, "Short-term Wind Power Forecasting Using Nonnegative Spare Coding," in *49th Annual Conference on Information Sciences and Systems (CISS)*, 2015, pp. 1-5.

[5] R. Tawn, J. Browell, and I. Dinwoodie, "Missing Data in Wind Farm Time Series: Properties and Effect On Forecasts," *Electric Power System Research*, p. 189, 2019.

[6] A. Okutan, G. Werner, S. J. Yang, and K. McConky, "Forecasting Cyberattacks With Incomplete, Imbalanced and Insignificant Data," *Cybersecurity*, vol. 1, no. 1, pp. 1-16, 2018.

[7] F. A. Adnan, M. H. Zakaria, and S. Ibrahim, "60-year Research History of Missing Data: A Bibliometric Review on Scopus Database (1960-2019)," *Applied Mathematics and Computational Intelligence*, vol. 9, no. 1, pp. 75-86, 2020.

[8] M. C. M. De Goeij, M. Van Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker, "Multiple Imputation: Dealing With Missing Data," *Nephrology Dialysis Transplantation*, vol. 28, no. 10, pp. 2415-2420, 2013.

[9] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, "A Review of Missing Values Handling Methods On Time-Series Data," in *2016 International Conference On Information Technology Systems and Innovation (ICITSI)*, 2016, pp. 1-6.

[10] M. K. Choong, M. Charbit, and H. Yan, "Autoregressive-model-based Missing Value Estimation for DNA Microarray Time Series Data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 1, pp. 131-137, 2009.

[11] K. Strike, K. E. Emam, and N. Madhavji, "Software Cost Estimation With Incomplete Data," *IEEE Transactions On Software Engineering*, vol. 27, no. 10, pp. 890-908, 2001.

[12] T. Moon, S. Hong, H. Y. Choi, D. H. Jung, S. H. Chang, and J. E. Son, "Interpolation of Greenhouse Environment Data Using Multilayer Perceptron," *Computers and Electronics in Agriculture*, vol. 166, p. 105023, 2019.

[13] S. Sun, Y. Bao, M. Lu, W. Liu, X. Xie, C. Wang, and W. Liu, "A Comparison of Models For The Short-term Prediction of Rice Stripe Virus Disease And Its Association With Biological and Meteorological Factors," *Shegtai Xuebo/Acta Ecologica Sinica*, vol. 36, no. 3, pp. 166-171, 2016.

[14] J. Liu, S. Kumar, and D. P. Palomar, "Parameter Estimation of Heavy-Tailed AR Model with Missing Data Via Stochastic EM," *IEEE Transactions on Signal Processing*, vol. 67, no. 8, pp. 2159-2172, 2019.

[15] H. Hassani and M. R. Yeganegi, "Selecting Optimal Lag Order in Ljung-Box Test," *Physica A: Statistical Mechanics and Its Applications*, vol. 541, p. 123700, 2020.