Research Journal of Applied Sciences, Engineering and Technology 6(10): 1780-1784, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 22, 2012 Accepted: December 20, 2012 Published: July 20, 2013

Application of the Single Imputation Method to Estimate Missing Wind Speed Data in Malaysia

[1]Nurulkamal Masseran, [1, 2]Ahmad Mahir Razali, [1, 2]Kamarulzaman Ibrahim,
[2, 3]Azami Zaharim and [2]Kamaruzzaman Sopian
[1]Centre for Modelling and Data Analysis (DELTA), School of Mathematical Sciences, Faculty of Science and Technology,
[2]Solar Energy Research Institute (SERI),
[3]Head of Project Group of Renewable Energy Resources Analysis, Policy and Energy Management, Renewable Energy Niche, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor D.E., Malaysia

Abstract: In almost all research fields, the procedure for handling missing values must be addressed before a detailed analysis can be made. Thus, a suitable method of imputation should be chosen to address the missing value problem. Wind speed has been found in engineering practice to be the most significant parameter in wind power. However, researchers are sometimes faced with the problem of missing wind speed data caused by equipment failure. In this study, we attempt to implement four types of single imputation methods to estimate the wind speed data from three adjacent stations in Malaysia. The methods, known as the site-dependent effect method, the hour mean method, the last and next method, and the row mean method, are compared based on the index of agreement to identify the best method for estimating the missing values. The results indicate that the last and next is the best of the three methods for estimating the missing data for the wind stations considered.

Keywords: Imputation technique, missing at random, site-dependent effect method, wind speed

INTRODUCTION

Missing data are a concern in almost all research fields and need to be addressed before data analysis. There are several reasons why wind speed data may be missing, including malfunctioning equipment, terrible weather, incorrect data recording, and so on. Wind speed data that are missing for these types of reasons can be categorized as Missing Completely at Random (MCAR) because their absence does not depend on other variables.

Among the previous studies on the missing data problem is the work by Plaia and Bondi (2006). They proposed a new single imputation method known as the Site-Dependent Effect Method (SDEM) and compared its performance to other single and multiple imputation methods. The SDEM method was compared with other imputation methods using the correlation coefficient, index of agreement, root mean square deviation and mean absolute deviation as measures of performance. The SDEM method was found to be the best of the methods compared in terms of all of the performance measures considered. Junninen *et al.* (2004) evaluated and compared univariate and multivariate methods for missing data imputation in air quality data sets. Among the univariate methods studied were linear interpolation, spline interpolation, and the nearest neighbor method, while the multivariate methods were regression-based imputation, the multivariate nearest neighbor method, the self-organizing map and multi-layer perception. The performance of each method was evaluated with respect to the index of agreement, the squared correlation coefficient, the root mean squared error and the mean absolute error with bootstrapped standard error. The results indicated that the univariate methods are dependent on the length of the gap in time in the data and that their performance depends on the variable under study. The results also showed that a slight improvement in the performance of multivariate methods can be achieved using hybridization, and a more substantial improvement can be achieved by using multiple imputation, where a final estimate is derived from the outputs of several multivariate fill-in methods.

MISSING DATA MECHANISM

Technically, missing data can be classified into three categories; Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). Consider a set of wind speed data $X = x_j$ and an indicator (dummy) variable $M = m_j$,

Corresponding Author: Ahmad Mahir Razali, Centre for Modelling and Data Analysis (DELTA), School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor D.E., Malaysia

where m_j has a value of 1 if M is missing and 0 if M is observed. The missing data mechanism is expressed by the conditional distribution of M given Y, say, f (M|X, θ), where θ denotes unknown parameters. Let X_{obs} and X_{miss} denote the observed and missing components of X, respectively. The missing data can be classified as MCAR if the following is true for all X:

$$f\left(M \mid X, \theta\right) = f\left(M \mid \theta\right) \tag{1}$$

If the following is true for all X_{miss} and θ:

$$f\left(M \mid X, \theta\right) = f\left(M \mid X_{obs}, \theta\right) \tag{2}$$

Then, the missing data are said to MAR. However, if the following is true:

$$f\left(M \mid X, \theta\right) = f\left(M \mid X_{miss}, \theta\right) \tag{3}$$

The missing data are said to be NMAR. In this study, as mentioned above, the wind speed data can be categorized as Missing Completely at Random (MCAR) because their absence does not depend on other variables.

METHODOLOGY

A number of methods are available in the literature to address the missing value problem (Ding and Ross, 2012; Ferrari *et al*., 2012; Jung *et al*., 2007; Templ *et al*., 2011). However, in this study, we focus on single imputation method to provide a satisfactory but simple way to impute missing wind speed data.

Site-Dependent Effect Method (SDEM): This single imputation method was proposed by Plaia and Bondi (2006). SDEM considers spatial and temporal information to impute the missing values in the data. Table 1 shows the data set structure for the 3 wind stations considered in this study. Let x_{swdh} be a generic element of the data set, where s refers to the wind stations (s = 1, 2 and 3), w refers to the week (w = 1, 2, 3, …, 53), d refers to the day of the week (d = 1, 2, …, 7) and h refers to the hour (h = 1, 2, 3, …24). The SDEM method explicitly considers a week effect, a day effect and an hour effect. The SDEM method in this study is given by the following:

$$\hat{x}_{swdh} = \bar{x}_{.wdh} + \frac{1}{2}\left(\bar{x}_{sw..} - \sum_{s=1}^{3}\frac{\bar{x}_{sw..}}{3}\right)$$

$$+\frac{1}{2}\left(\bar{x}_{s.d.} - \sum_{s=1}^{3}\frac{\bar{x}_{s.d.}}{3}\right) + \frac{1}{2}\left(\bar{x}_{s..h} - \sum_{s=1}^{3}\frac{\bar{x}_{s..h}}{3}\right), \tag{4}$$

Table 1: Data set structure

Week of the year	Day of the week	Hour of the day	St 1	St 2	St 3
1	1	1	4.9	4.7	5
1	1	2	4.9	4.3	6
1	1	3	3.7	5.6	6.2
1	1	4	1.7	6.2	5.6
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
53	7	24	2.8	1.1	2.4

where,

$\bar{x}_{.wdh}$ = The mean of the values observed for all stations in week w, day d and hour h

$\bar{x}_{sw..}$ = The mean of the values observed in week w at site s

$\bar{x}_{s.d.}$ = The mean of the values observed at site s on day d

$\bar{x}_{s..h}$ = The mean of the values observed at site s in hour h

The SDEM method incorporates the spatial and temporal information from each station involved in this study.

Hour Mean Method (HMM): The hour mean method imputes missing data from a given station using hourly information from the same station. According to Li *et al.* (1999), HMM fills in missing hourly observations by computing the mean for all known hourly observations for the same station at the same hour over the whole year. The HMM is given by the following:

$$\hat{x}_{swdh} = \bar{x}_{s..h}, \tag{5}$$

where,

$\bar{x}_{s..h}$ = The mean of the values observed at site s in hour

s = 1, 2 and 3

h = 1, 2,…, 24

Last and Next Method (LNM): The last and next method imputes missing values in the data from a given station by incorporating information from the same station. LNM is performing by assigning the average of the last known and next known observations to the missing value. LNM is given by the following:

$$\hat{x}_{swdh} = \frac{x_{swd(h-1)} + x_{swd(h+1)}}{2} \tag{6}$$

where,

s = 1, 2 and 3

w = 1, 2,…, 53

d = 1, 2,…, 7

h = 1, 2,…, 24

However, LNM can only be applied for a single missing value at a time. For cases involving of missing values, such as values with a gap length l≥2, LNM is

not applicable. We suggest a more generalized way to formulate LNM to overcome its limitations. LNM can also be performed by assigning the average of the last known and next known observations to the missing value of a particular day, which may be written as follows:

$$\hat{x}_{swdh} = \frac{x_{sw(d-1)h} + x_{sw(d+1)h}}{2} \quad (7)$$

where,
s = 1, 2 and 3
w = 1, 2,…, 53
d = 1, 2,…, 7
h = 1, 2,…, 24

In addition, LNM can be performed by assigning the average of the last known and next known observations to the missing value of a particular week, which may be written as follows:

$$\hat{x}_{swdh} = \frac{x_{s(w-1)dh} + x_{s(w+1)dh}}{2} \quad (8)$$

where,
s = 1, 2 and 3
w = 1, 2,…, 53
d = 1, 2,…, 7
h = 1, 2,…, 24

With these formulations, LNM is more generalized and can easily be used to impute missing values with long gap lengths. However, care should be taken to examine the weekly and daily patterns in the data before applying this method (Plaia and Bondi, 2006; Engels and Diehr, 2003).

Row Mean Method (RMM): The row mean method imputes missing data from a given station using hourly information from the same station. Row mean method uses hourly information from the same station in order to impute the missing data. RMM is performed by computing the mean of all known observations in the same row of the data matrix as the missing data. HMM is given by the following:

$$\hat{x}_{swdh} = \overline{x}_{.wdh} \quad (9)$$

where,
w = 1, 2,…, 53
d = 1, 2,…, 7
h = 1, 2,…, 24

Index of agreement as a performance indicator: The index of agreement has been used to evaluate the effectiveness of each imputation method. Let O_i = The i^{th} data point $\bar{O}$ = The average of the observed data, let

Table 2: Descriptive statistic of missing data for each station

Station	% missing	% gap length (*l*)			
		l = 1	1<*l*≤3	3<*l*≤12	*l*>12
K. Terengganu (KT)	1.052	66.67	22.22	11.11	0
Kertih (KR)	14.566	6.67	13.33	46.67	33.33
Kemaman (KM)	2.564	33.33	33.33	16.67	16.67

P_i denote the i^{th} imputed data point and $\bar{P}$ denote the average of the imputed data, and let N denote the number of imputation. Thus, the index of agreement is given by the following:

$$d = \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left(|P_i - \bar{O}| + |O_i - \bar{O}| \right)^2} \right] \quad (10)$$

The index of agreement evaluates the effectiveness of each imputation method by measuring the discrepancy between observed and imputed values of missing data. The smaller the index value, the better the imputation of missing data.

ANALYSIS, RESULTS AND CONCLUSION

Before a detailed analysis is made, it is important to explore some descriptive statistics to obtain some preliminary information about the data. Table 2 shows a certain percentage of missing data for each station. K. Terengganu and Kemaman stations have a small percentage of missing data, approximately 1.062 and 2.564%, respectively. However, Kertih station has quite a large percentage of missing data, approximately 14.566%. To examine these gaps in detail, we use the four gap length (*l*) categorizes identified by Plaia and Bondi (2006) for missing data, namely 1 observation gap, 1 to 3 observation gaps, 3 to 12 observation gaps and more than 12 observation gaps. From Table 2, we can see the distribution of missing values for each station according to this categorization. Figure 1 shows the frequency distribution of the gap length for each station.

To apply the imputation methods described above, especially the SDEM method, it is informative for us to examine the spatial and temporal characteristics of the data from each station. Figure 2 shows a line plot for the mean values observed on *w* week for each station. It is found that week mean plots for K. Terengganu and Kertih station indicate quite similar trends, while the week mean plot for Kemaman station has a different trend. However, the difference in the Kemaman trend is not significant. This indicates that there is some correlation in the week effect for each station. Figure 3 and 4 show the line plots for the mean values observed on *h* hour and *d* day.

The hour mean values and day mean values exhibit approximately similar patterns for each station except that they differ by a fairly constant amount. The similarity of the patterns for each station indicate the

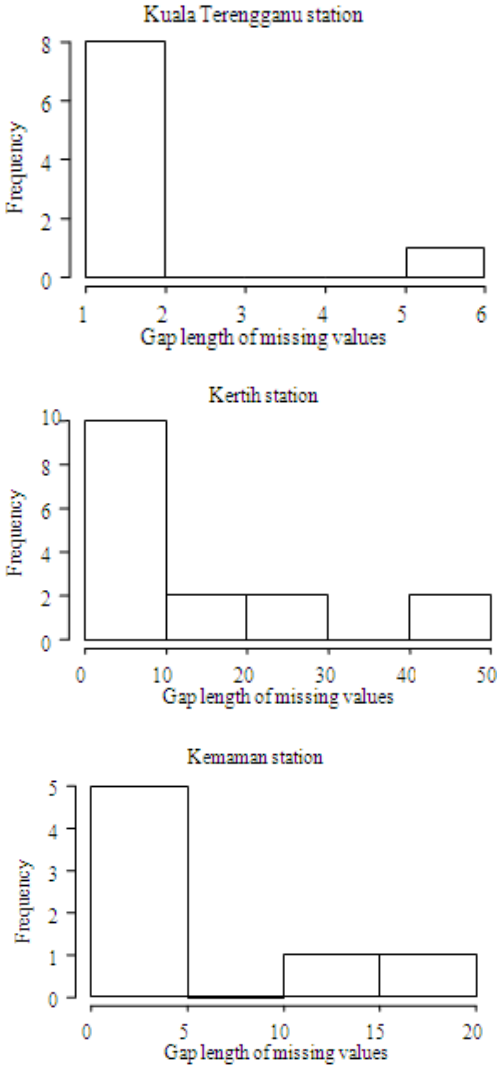

Fig. 1: The frequency distribution of the gap length for each station

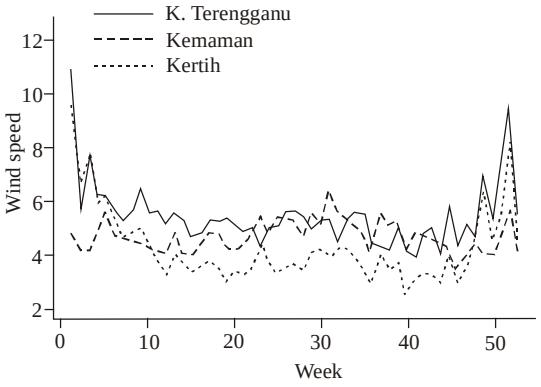

Fig. 2: A line plot for the mean values observed for *w* week for each station

presence of some valuable information that can be used to estimate the missing values, particularly with the SDEM method.

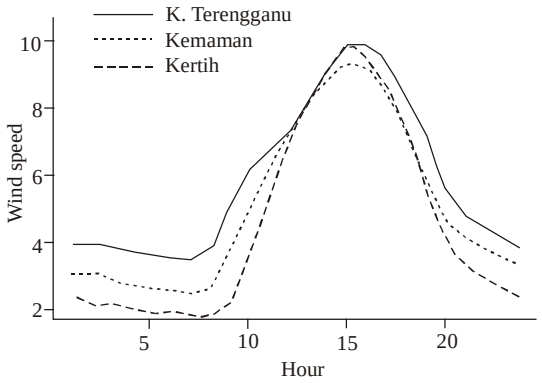

Fig. 3: A line plot for the mean values observed for *h* hour for each station

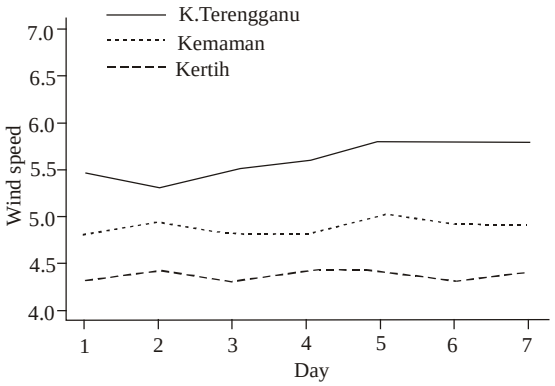

Fig. 4: A line plot for the mean values observed for *d* day for each station

Table 3: The index of agreement for each imputation method

	Index of agreement		
	Wind station		
Method	KT	KR	KM
SDEM	0.2982	0.2803	0.2560
HMM	0.3462	0.4629	0.1867
LNM	0.1803	0.1439	0.0472
RMM	0.3170	0.2649	0.3905

Because we already know that there is some correlation in the wind speed data from the three stations, based on our subjective evaluation of Fig. 1 to 3, it is informative for us to apply each imputation method, particularly SDEM, to address the missing data problem. To determine which method provides the best imputation of the missing values, the index of agreement is used as the measure of performance. The evaluation process begins by simulating the incomplete data set from the portion of the original data with no missing values. Each imputation method is then applied to determine the simulated missing value from the simulated data set. The performance of each method is calculated using the index of agreement. The method with the smallest index value is considered the best method to estimate our missing wind speed data. Table 3 shows the index of agreement results for each

method. For the data considered in this study, we found that the last and next method had the smallest index of agreement values for all of the stations, which indicates that the last and next method is the best method for estimating the missing data for the stations considered in this study. The site-dependent effect method is found to be the second-best method for estimating the missing value. Thus, we conclude here that the last and next method is the best method for estimating missing values in the data used in this study.

ACKNOWLEDGMENT

The authors are indebted to the staff of the Department of Environment for providing data for use in this study. This research would not have been possible without the sponsorship of the Universiti Kebangsaan Malaysia and Ministry of Higher Education, Malaysia (grant number UKM-GUP-2011-213 and OUP-2012-064).

REFERENCES

Ding, Y. and A. Ross, 2012. A comparison of imputation methods for handling missing scores in biometric fusion. Pattern. Recogn., 45: 919-933.

Engels, J.M. and P. Diehr, 2003. Imputation of missing longitudinal data: A comparison of methods. J. Clin. Epidemiol., 56: 968-976.

Ferrari, P.A., P. Annoni, A. Barbiero and G. Manzi, 2012. An imputation method for categorical variables with application to nonlinear principle component analysis. Comput. Stat. Data An., 55: 2410-2420.

Jung, H.Y., Y.J. Park, Y.J. Kim, J.S. Park, K. Kimm and I. Koh, 2007. New methods for imputation of missing genotype using linkage disequilibrium and haplotype information. Inform. Sci., 177: 804-814.

Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen, 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ., 38: 2895-2907.

Li, K.H., N.D. Le, L. Sun and V.J. Zidek, 1999. Spatial-temporal model for ambient hourly PM_{10} in vancouver. Environmetrics, 10: 321-328.

Plaia, A. and A.L. Bondi, 2006. Single imputation method of missing data in environmental pollution data sets. Atmos. Environ., 40: 7316-7330.

Templ, M., A. Kowarik and P. Filzmozer, 2011. Iterative stepwise regression imputation using standard and robust methods. Comput. Stat. Data An., 55: 2793-2806.