Imputation of Missing Values for Solar Irradiance Data under Different Weathers using Univariate Methods

Noor Bariah Mohamad¹, Boon-Han Lim*,¹,²,³, An-Chow Lai¹

¹ Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Bandar Sungai Long, 43000 Kajang, Selangor, Malaysia
² Centre for Photonics and Advanced Materials Research, Universiti Tunku Abdul Rahman, Malaysia.
³ Centre for Powers and Electricity, Universiti Tunku Abdul Rahman, Malaysia.
E-mail: limbhan@utar.edu.my

Abstract. Significant investment risks of large photovoltaic (PV) systems are uncertainties of energy yield predictions from a PV power plant. Unfortunately, solar irradiance weather datasets often have missing values due to operational issues; hence techniques of imputing or recovering the missing values become essential. This paper examined five statistical imputation methods that are frequently adopted in missing values analysis for daily solar irradiance series based on two different weather conditions in the tropical climate with 10% to 50% of missing values and the methods were evaluated using performance indicators such as normalised root-mean-square at standard test conditions (nRMSESTC). The results show that during Sunny weather, the Bezier curve and Stineman interpolation give the lowest nRMSESTC for 10% and 20% to 50% of daily missing values, respectively. Meanwhile, during Largely Cloudy weather, three methods share the best estimations, which are Linear interpolation for 10% missing values, Stineman interpolation for 20%, 30% and 50% missing values, and Spline interpolation for 40% of missing values.

1. Introduction

The photovoltaic (PV) industry is growing rapidly worldwide due to its environmental-friendly technology. When designing or operating a PV system, the availability of solar irradiance (W/m²) information is essential as it affects the bankability of a proposed project or the healthiness of an operating project. Furthermore, the information has to be complete, usually on an annual basis, for achieving an optimal design. The ground-mounted measurement of solar irradiance is well known as a more accurate measurement method. However, failures in the measurement process, such as cable error, malfunction of the data logger, and unnecessary man-make shading, will lead to incomplete solar irradiance data. Therefore, a method to resolve the missing data by imputation is necessary. Generally, the imputation studies can be divided into two main approaches, i.e., machine learning and statistical methods. Machine learning methods as proposed by Ref. [1] and Ref. [2] require more than one climatology parameter such as cloudiness, precipitation, temperature, etc., as inputs. Contrarily, the statistical method offers an advantage that allows imputation to be carried out with only one parameter. Often, it is costly to equip a weather station with all the necessary toolsets to collect a complete set of data containing other climatology parameters. In practice, most PV projects are only installed with a solar irradiance meter.

Furthermore, solar irradiance is highly dependent on the state of the sky. Das et al. [3] documented inadequate output on rainy days relative to sunny days for PV power generation prediction studies. The result indicates that a different type of weather requires a different imputation method. Therefore, this study investigates five frequently used statistical methods to impute the missing data for daily
solar irradiance for two weather conditions in the tropical climate with various percentages of missing values.

2. Methodology

2.1. Data

The global horizontal solar irradiance (GHI) datasets were measured at a 5-minute interval using a pyranometer located at the roof of the highest building in the area of the Universiti Tunku Abdul Rahman (UTAR) Sungai Long Campus, Kajang, Malaysia. Two types of weather, i.e., Sunny and Largely Cloudy, have been chosen for this study. Figures 1 and 2 show the daily GHI profiles during the Sunny and Largely Cloudy weather. Following that, 10%, 20%, 30%, 40%, and 50% of the daily data were randomly removed from each of the complete datasets to form 10 sets of incomplete datasets for our studies to evaluate the accuracy of imputation methods. The complete datasets act as the reference for error calculation.

2.2. Univariate Methods

Univariate imputation estimates one value for each missing datum. Five univariate imputation methods, which are linear interpolation (LI), spline interpolation (SpI), Stineman interpolation (StI), Bezier curve, and simple moving average (SMA) were applied to estimate the missing values.

2.2.1. Interpolations. Noted that, y is the interpolant, x is the time point of the interpolant, (x_i, y_i) and (x_{i+n}, y_{i+n}) are coordinates of the start points and endpoints with (i = 0, 1, 2, ...) and (n = 1, 2, ...).

LI is a baseline piecewise, fits a straight line between the endpoints of the gap, defined as [4]:

\[ y = y_i + \frac{(y_{i+n} - y_i)}{(x_{i+n} - x_i)}(x - x_i) \]  \hspace{1cm} (1)

SpI fits quadratic piecewise polynomials to a series of observed data points, given by [4]:

\[ y = A + b_2(x - x_i)(x - x_{i+1}) \]  \hspace{1cm} (2)

where A is from Equation (1), and 

\[ b_2 = \frac{[(y_{i+n+1} - y_{i+n})/(x_{i+n+1} - x_{i+n}) - (y_{i+1} - y_i)/(x_{i+1} - x_i)]}{(x_{i+1} - x_i)}. \]

In StI, the interpolant point is calculated by two conditions [5], as follows:

\[ y = y_i + \frac{[\Delta y(\Delta y_{i+n})/\Delta y_{i+n}]}{[\Delta y_{i+n}]} \]  \hspace{1cm} (3)

if \( \Delta y_{i+n} > 0 \)

\[ y = y_i + \frac{[\Delta y(\Delta y_{i+n})(2x - x_{i+n} - x_i)]/[(\Delta y_{i+n})(x_{i+n} - x_i)]}{\Delta y_{i+n}} \]  \hspace{1cm} (4)

if \( \Delta y_{i+n} < 0 \)

where \( \Delta y_i \) and \( \Delta y_{i+n} \) are verticals distance from the point \((x_i,y_i)\) to a line through \((x_{i+1},y_{i+1})\) with slope \( y_{i+1}' \) by \( \Delta y_{i+1} = y_{i+1}'(x - x_i) \) for i and i+n. Point \( y_0 \) is determined by calculating the ordinate corresponding to time \( x \), by \( y_0 = y_i + s_i(x - x_i) \). \( s_i \) is the slope of the line connecting \((x_i,y_i)\) and \((x_{i+n},y_{i+n})\), by \( s_i = (y_{i+n} - y_i)/(x_{i+n} - x_i) \). The last step is estimating the slope \( y_{i+1}' \). It is determined by letting consecutive points of I,
J, and K satisfying either slope(\(IJ\)) > \(y_i'\) > slope(\(JK\)) or slope(\(IJ\)) < \(y_i'\) < slope(\(JK\)). Thus, the slope \(y_i'\) is given as:

\[
y_i' = \frac{(y_j - y_i)(x_k - x_j)^2 + (y_k - y_j)(x_j - x_i)^2 + (y_j - y_k)(x_k - x_i)^2}{(x_j - x_i)[(x_k - x_j)^2 + (y_k - y_j)^2][x_j - x_i][x_k - x_j]^2}
\]

### 2.2.2. Parametric Curve
The degree three Bezier curve required four points, referred as \(b_0\) and \(b_1\) be the endpoints and \(b_1\) and \(b_2\) be the inner points. The inner points are used to control curve development. The general form of the Bezier curve given as [6]:

\[B(t) = \sum b_i B_i^3(t), \text{ where } 0 \leq t \leq 1\]

where \(b_i\) is the Bezier control points, and \(B_i^3(t)\) equal to \([3!/(i!)(3-i)!]\) \((1-t)^3+i\). The inner points, \(b_1\) and \(b_2\), can be calculated through the tangents between the end points, \(b_0\) and \(b_3\), respectively, as follows:

\[b_1 = b_0 + (1/3)(x_2-x_1)F'(x_1), \quad b_2 = b_3 + (1/3)(x_2-x_1)F'(x_2)\]

where \(F'(x_1)\) and \(F'(x_2)\) are the estimated slopes at the start and the end of a 5-min interval of solar irradiance.

### 2.2.3. Simple Moving Average
The concept of SMA is by averaging the \(k\) observation in replacing each of the missing values, known as the window size. In our study, we only consider \(k = 2\). Equation (8) gives the formula of SMA:

\[\text{SMA}^k_i = \frac{(A_{i-1} + A_{i-2} + \ldots + A_{i+k})}{k}\]

where, \(\text{SMA}^k_i\) is the simple moving average for window \(k\) included in \(t\) period, \(A_{i-1}\) to \(A_{i+k}\) are the solar irradiance number involved in a period of \(t-1\) to \(t-k\), while \(k\) is the number of periods in average.

### 2.3. Performance Indicators
The performances of the five imputation methods were evaluated using the following common measures: Mean Absolute Error (MAE) and Normalized Root Mean Square (nRMSE) under standard test condition (STC), respectively, as follows [7]:

\[\text{nRMSE}_{\text{STC}} = \frac{[\sum_{i=1}^{N} (J(t) - \hat{J}(t))^2]^{1/2}}{1000}\]

\[\text{MAE} = \frac{\sum_{i=1}^{N} |J(t) - \hat{J}(t)|}{N} \]

where \(J(t)\) is the original measured GHI data, \(\hat{J}(t)\) is the predicted GHI data based on the imputation methods, \(N\) is the number of the data, and 1000 W/m² is the solar irradiance value used for STC. The observed and predicted data were compared to select the best method for estimating missing values.

### 3. Results and Discussions

#### 3.1. Solar Irradiance Series for Sunny Weather
Figure 3 shows the bar chart of nRMSE_{STC} and MAE for the five imputation methods based on the solar irradiance series for Sunny weather for five different settings of missing data, i.e., 10%, 20%, 30%, 40%, and 50% of the datasets are missing values. For dataset less than 30% missing values, all methods except SMA can impute the missing data with nRMSE_{STC} less than 0.013 and MAE less than 5. The dataset with 10% missing values exhibits the smallest error, for both nRMSE_{STC} and MAE, for all methods. We observed that the errors increase as the percentage of missing values increases. This finding is supported by Paul et al. [8] that the number of missing values significantly affects the efficiency of imputations.
Among the univariate methods, the Bezier curve method gives the lowest nRMSE_{STC} and MAE for the 10% missing values. Whereas, for 20% to 50% missing values, StI performed the best. StI has the advantage over other methods as it is robustly sustained the monotonicity of the interpolated period [9]. Meanwhile, an SMA gives the highest errors for all percentages of missing values.

**Figure 3.** Comparison of imputation methods of solar irradiance for Sunny weather with 10% to 50% missing values (a) nRMSE_{STC} (b) MAE

3.2. Solar Irradiance Series for Largely Cloudy Weather

Figure 4 shows the bar chart of nRMSE_{STC} and MAE for the same five methods based on solar irradiance series for Largely Cloudy weather for 10% to 50% missing values. StI, LI, and SpI give nRMSE_{STC} and MAE less than 0.01 and 5 respectively for all percentages of missing values. LI gave the best estimation for 10% missing values of solar irradiance in both nRMSE_{STC} and MAE values. As the percentage of missing values increases to 20% and 30%, StI is observed with the smallest errors. Both LI and StI perform well in interpolating smooth curves. However, StI is more robust if the solar irradiance data poses abrupt slope changes. In the case of 40% and 50% missing values, SpI and StI lead in the performance indicators, respectively.

**Figure 4.** Comparison of imputation methods of solar irradiance for Largely Cloudy weather with 10% to 50% missing values (a) nRMSE_{STC} (b) MAE.

The best methods for each percentage of missing values for the two types of weather are summarised in Table 1. It indicates that different methods should be adopted when imputing different percentages of missing values as well as for different weather, even though StI can be chosen most of the time. However, in this study, only a one-day solar irradiance series was used for each weather. It could be the reason at 40% missing values on Largely Cloudy weather, the best method is changed to SpI and then returned to StI at 50% missing values. Therefore, further study can be performed for many days to obtain more conclusive results.
Table 1. The method with the smallest nRMSE_{STC} values at a specific percentage of missing values for Sunny and Largely Cloudy weathers.

| Weathers          | Imputation Methods |
|-------------------|--------------------|
|                   | Bezier Curve       | StI | LI    | SpI   | SMA    |
| Sunny             | 10%                | 20%, 30%, 40%, 50% |
| Largely Cloudy    | 20%, 30%, 50%      | 10% | 40%   |

4. Conclusion
In this study, we examined five univariate imputation methods to fill in missing values at different percentages during Sunny and Largely Cloudy weathers, and then its performances were analyzed. Amongst five imputation methods, StI yields the lowest nRMSE_{STC} and MAE values for 20% to 50% missing values for Sunny weather as well as 20%, 30%, and 50% missing values for Largely Cloudy weather. Meanwhile, the Bezier curve and LI yield the lowest errors at 10% missing values for Sunny and Largely Cloudy weather respectively. Exceptionally, SpI performs the best for 40% missing values of Largely Cloudy weather.

5. References
[1] J. Park and R. Engineering, “Comparison of spatial interpolation methods for estimating solar radiation in South Korea,” AWER Procedia Inf. Technol. Comput. Sci., vol. 04, no. 1, pp. 608–614, 2013.
[2] T. Kim, W. Ko, and J. Kim, “Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting,” Appl. Sci., vol. 9, no. 1, pp. 1–18, 2019.
[3] U. K. Das et al., “SVR-based model to forecast PV power generation under different weather conditions,” Energies, vol. 10, no. 7, pp. 1–17, 2017.
[4] M. Noor, Y. Ahmad, R. Nor, A. Mohd, and M. Al, “Estimation of missing values in air pollution data using single imputation techniques,” Sci. Asia, vol. 34, pp. 341–345, 2008.
[5] G. M. E. Perillo and M. C. Piccolo, “An interpolation method for estuarine and oceanographic data,” Comput. Geosci., vol. 17, no. 6, pp. 813–820, 1991.
[6] A. Saaban, L. Zainudin, and M. N. A. Bakar, “On piecewise interpolation techniques for estimating solar radiation missing values in Kedah,” in AIP Conference Proceedings, 2014, pp. 217–221.
[7] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, “Methods for imputation of missing values in air quality data sets,” Atmos. Environ., vol. 38, no. 18, pp. 2895–2907, 2004.
[8] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, “The proportion of missing data should not be used to guide decisions on multiple imputation,” J. Clin. Epidemiol., vol. 110, pp. 63–73, 2019.
[9] S. P. Roy, P. G. Arias, V. R. Lecoustre, D. C. Haworth, H. G. Im, and A. Trouvé, “Development of high fidelity soot aerosol dynamics models using method of moments with interpolative closure,” Aerosol Sci. Technol., vol. 48, no. 4, pp. 379–391, 2014.

Acknowledgements
The research was supported by the Ministry of Higher Education Malaysia through Fundamental Research Grant Scheme (FRGS/1/2018/TK10/UTAR/02/10), Universiti Tunku Abdul Rahman through UTAR Research Fund (IPSR/RMC/UTARRF/2018-C2/L04), and Malaysian Photovoltaic Industry Association’s crowdfunding (Memorandum of Agreement dated 29th July 2019). The authors would like to express sincere gratitude to the three parties for financial support. Besides, the authors would like to thank Chong Kok Keong for sharing the pyranometer for solar irradiance measurement.