Inference of missing data in photovoltaic monitoring datasets

This item was submitted to Loughborough University’s Institutional Repository by the/an author.

Citation: KOUBLI, E. ...et al., 2016. Inference of missing data in photovoltaic monitoring datasets. IET Renewable Power Generation, 10(4), pp.434-439.

Additional Information:

- This is an Open Access Article. It is published by IET publishing under the Creative Commons Attribution 3.0 Unported Licence (CC BY). Full details of this licence are available at: http://creativecommons.org/licenses/by/3.0/

Metadata Record: https://dspace.lboro.ac.uk/2134/20351

Version: Published

Publisher: © The Authors. Published by Institution of Engineering and Technology (IET)

Rights: This work is made available according to the conditions of the Creative Commons Attribution 3.0 Unported (CC BY 3.0) licence. Full details of this licence are available at: http://creativecommons.org/licenses/by/3.0/

Please cite the published version.
Inference of missing data in photovoltaic monitoring datasets

Eleni Koubli, Diane Palmer, Paul Rowley, Ralph Gottschalg

Centre for Renewable Energy Systems Technology (CREST), School of Electronic, Electrical and Systems Engineering, Loughborough University, Loughborough, LE11 3TU, UK
✉ E-mail: E.Koumpli@lboro.ac.uk

Abstract: Photovoltaic (PV) systems are frequently covered by performance guarantees, which are often based on attaining a certain performance ratio (PR). Climatic and electrical data are collected on site to verify that these guarantees are met or that the systems are working well. However, in-field data acquisition commonly suffers from data loss, sometimes for prolonged periods of time, making this assessment impossible or at the very best introducing significant uncertainties. This study presents a method to mitigate this issue based on back-filling missing data. Typical cases of data loss are considered and a method to infer this is presented and validated. Synthetic performance data is generated based on interpolated environmental data and a trained empirical electrical model. A missing month of energy readings has been replenished, reproducing PR with an average daily and monthly mean bias error of about −1 and −0.02%, respectively, for a crystalline silicon module. The PR is a key property which is required for the warranty verification, and the proposed method yields reliable results in order to achieve this.

1 Introduction

The number of photovoltaic (PV) installations in the UK has increased from a few tens of MWp in 2010 to more than 6 GWp in June 2015 [1], indicating that PV is a rapidly growing industry. The majority of these systems will operate as financial investments, and in order to manage investment risk associated with system yield shortfalls, many larger scale PV systems are covered by energy performance guarantees. These are commonly based on attaining a specified performance ratio (PR), as this allows for location-specific variations in meteorological conditions. Energy yield, and to a lesser extent, PR are key performance indicators for owners, investors and operators and they can only be obtained through effective monitoring of PV systems, as highlighted in several studies [2–5]. However, due to malfunctions such as power outages, communication failures or component faults, data may be incomplete. This will affect, as shown below, the PR calculated for the system and may hide incidents that would trigger warranty cases or cause unnecessary warranty claims. There are no validated strategies that deal with this issue, particularly in maritime climates such as the UK, and any attempts to backfill data using previous dates or days from previous years are at best temporary with very high uncertainty attached to these methods. This is because such strategies do not take into account variable weather systems or PV component degradation. Utilising average values from dates close to the missing period may give an estimation of PR, but such methods are not adequate to estimate long-term energy yields of the system, especially when missing periods are extended from a few weeks to even months. Therefore, an issue remains of how to back-fill lost data values, and to arrive at a valid monthly or annual PR which is required in order to verify the warranties or to predict return on investment.

2 Methodology development and validation

Given the aim is to back-fill missing periods appropriately, then in order to infer data reliably, actual weather patterns as well as specific system performance need to be taken into account. Thus, it is not sufficient to just replace missing data with previous data, or any alternative method that does not consider actual meteorological conditions as this may introduce significant errors. Rather, the method presented here considers local weather phenomena and their relationship to performance variations of the actual PV systems. This requires two elements, assessing meteorological as well as electrical data as both systems may fail independently. Meteorological data for the missing period is obtained here from interpolating from a network of about 80 meteorological monitoring stations operated by the UK met-office [6]. The local irradiance is calculated from interpolating these using Kriging [7] and correcting the horizontal irradiance to the site installation by employing separation, into beam and diffuse, and translation, into plane of the array, algorithms. The same interpolation technique is applied for ambient temperature. Ambient temperature is corrected to module temperature by employing a simple thermal model. The method is described in more detail in Sections 2.1 and 2.2.

The energy output of missing periods is estimated using an empirical electrical model. The underlying coefficients are obtained by ‘training’ the model with the available past data (see Section 2.3), i.e. determining system specific characteristics. The validity of the proposed method is assessed using the following metrics:

(i) Root-mean-square error (RMSE)
(ii) Mean absolute error (MAE)
(iii) Mean bias error (MBE)

The RMSE describes the random error in a distribution and tends to increase with outliers, MAE describes the absolute error and MBE indicates whether the model overestimates or underestimates the measurement value, which is also expressed as the ‘systematic’ error of the distribution. MBEs close to zero signify an unbiased distribution. It should be noted that, although the word ‘error’ is commonly used in statistical analysis, ‘difference’ would here be
more appropriate since the true values are not actually known, as sensor uncertainty is not taken into account.

Real measurements of in-plane irradiation, ambient and module temperature, as well as energy output were used for the validation. The datasets used were specifically from two PV modules from CREST’s outdoor monitoring system (COMS3) [8] whose properties are listed in Table 1.

### 2.1 Irradiance and temperature interpolation

The analysis is based on meteorological data, namely global horizontal irradiation and ambient temperature, acquired from more than 80 ground meteorological stations on a national scale through Met Office Integrated Data Archive System (MIDAS) [6]. The PR of a PV system depends on the irradiation (H in Wh/m²) received according to (1)

\[
PR = \frac{E \cdot G_{STC}}{H \cdot P_{STC}}
\]

where \(E\) is the energy output (Wh), \(P_{STC}\) (W) and \(G_{STC}\) (W/m²) are the nominal power of the system and irradiance at Standard Testing Conditions (STC), respectively.

Horizontal irradiance is interpolated to a grid of points. The nearest point of the PV system is selected and separation (Ridley et al. [9]) and translation algorithms (Hay et al. [10] with Reindl correction [11]) are employed to calculate in-plane irradiance given that the location, orientation and tilt of the system are known. The specific separation model was chosen based on empirical observations, which demonstrated that it delivered the best results in comparison with several separation algorithms for UK [7]. Similarly, a previous work in Loughborough [12] has shown that the all-sky model by Reindl et al. delivers the best results for the UK climate. Both the horizontal irradiation and the ambient temperature data are interpolated using Kriging, which has been proven to perform well in comparison with various climate data interpolation methods [7, 13].

### 2.2 Thermal and electrical model

Module temperature is calculated from in-plane irradiance and ambient temperature using the thermal model presented in [14]

\[
T_m = T_a + k \cdot G
\]

where \(T_m\) (K), \(T_a\) (K) and \(G\) (W/m²), are module temperature, ambient temperature and in-plane irradiance, respectively. The \(k\), or Ross coefficient is the modified thermal resistance of the module, modified in terms of influence of the mounting configuration of the array [15] a typical value of which is 0.02 (K·m²/W) for free-standing modules [16]. Ross’s model is a good choice in cases where irradiance and ambient temperature are the only available weather data. Furthermore, \(k\) can be readily obtained using outdoor measurements of module and ambient temperature and horizontal irradiance. In this work, \(k\) was obtained experimentally for each module by linear fitting of \((T_m-T_a)\) against \(G\) for one year’s worth of data. Equation (2) was used by taking the hourly values of irradiance as proposed in [17].

The electrical model was chosen based on both the available input data and its training capability. The chosen electrical model plays the role of the ‘learning machine’. It is based on a simplified King’s model for the maximum power point [18] and the formula used here has the following form [19]

\[
P(G, T) = G \cdot (1 + k_1 \ln(G) + k_2 \ln(G)^2 + k_3 T_m + k_4 T^*_m \ln(G) + k_5 T_m \ln(G)^2 + k_6 T^*_m)
\]

where \(P = P_{MP} / P_{STC}\), \(G = G / G_{STC}\) and \(T_m = T_{MP} - T_{STC}\) (STC = Standard Testing Conditions) and \(P_{MP}\) is the maximum power. The model yields a ‘3D power surface’. This model has been compared with a number of other models [20] where it was found that it performed well for a range of PV module technologies, on predicting annual energy output and it is also possible to combine measured data from many PV modules to obtain a general model for a given PV technology [19]. By changing the \(P_{STC}\) (3) can be used to describe an entire PV system. In this study, defining the coefficients for a specific system is essentially training the model based on the specific system characteristics and re-using the coefficients to predict the output of the missing period. Energy generation is calculated using sums of hourly averaged maximum power output. For the training process, past data are fed into the model and the coefficients \((k_1-k_6)\) are determined by means of a Marquardt–Levenberg optimisation algorithm [21]. To assess the reliability of the results, data quality checks and a training algorithm were used to determine the optimum training set for the model.

### 2.3 Extraction of the fitting coefficients

Quality checks are a critical step in every data analysis, the aim of which is to identify outliers that could corrupt the training process. Thus, the quality checks applied here focused on module temperature, irradiation and maximum power output. Graphical representation of the above parameters as well as logical controls based on (1) and (2) were employed according to the analyses described in [22, 23]. Only unshaded and fault-free systems were considered for this study.

The training algorithm is key for the acquisition of the optimum system’s coefficients, as it has been shown that device-specific characteristics, and thus implicitly the system characteristics, are one of the two uncertainties determining the model accuracy [20]. The requirements for the training need to be determined in terms of optimal training set’s size and how recent it should be with respect to the missing period of data, in order to achieve a maximum agreement. System performance is affected by both meteorological seasonal variations as well as by module technology [24, 25]. This seasonal nature of system performance is expected to have an impact upon the determination of the optimum training set. This study concluded that in all cases, the training set should maintain close proximity to the missing period, as this provides a better fit. More specifically, an average of 20 days before and 20 days after the missing set was found to be a sufficient data pool for the particular location, regardless of the position of the missing set throughout the year (i.e. during summer or winter etc.).

The training algorithm was applied separately for each one of the PV modules in Table 1. For the training process, past data were analysed using (3). The optimal training set was chosen based upon the lowest RMSE achieved and \(R^2\) values, in combination with the size of the training set. The training size was defined as the number of days before (noted as negative, going backwards in time) and after (noted as positive, going forwards in time) the missing period. The input training data were hourly measurements of power, module temperature and in-plane irradiance. Here, 1

---

**Table 1** Table of PV modules

<table>
<thead>
<tr>
<th>Name</th>
<th>Module type</th>
<th>Tilt angle, °</th>
<th>Orientation</th>
<th>Nominal power, W</th>
<th>Mounting type</th>
<th>Data origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>module A</td>
<td>crystalline silicon (c-Si)</td>
<td>32.5</td>
<td>south</td>
<td>245.0</td>
<td>open-rack</td>
<td>CREST outdoor monitoring system</td>
</tr>
<tr>
<td>module B</td>
<td>polycrystalline silicon (pc-Si)</td>
<td>32.5</td>
<td>south</td>
<td>245.0</td>
<td>open-rack</td>
<td>CREST outdoor monitoring system</td>
</tr>
</tbody>
</table>
month of missing data was considered (June 2014). This period was removed from the training set and was used as the validation set for the training algorithm.

It can be seen in Fig. 1a that although RMSE does not vary significantly across different training sets, there is a specific (red) area where it showed its lowest values. This area includes points that are closer to the missing period, which can be justified as seasonal dependence. The results also showed that very small training sets yielded the highest RMSE, e.g. using only several days before and after the missing period was not a sufficient data pool. This seems to be due to location and the local weather phenomena, whereas smaller training sets, are expected to suffice for less variable weather systems (for example, a Mediterranean summer). The choice of the training set’s size plays an important role as it must be large enough in order to obtain the optimal model coefficients which are then used to predict the energy output for the missing period.

The training algorithm defined the best set of coefficients \((k_1-k_6)\) which then provided the power surface shown in Fig. 1b. The optimal training size for this case was found to be 41 days in total.

Finally, the training process yields valid results if no significant changes (i.e. component failures) have occurred in the PV system during its operation while no data are available (i.e. during the missing period).

### 3 Results and discussion

#### 3.1 Analysis of interpolated climatic data

The following analysis is carried out for 1 year (2014). Irradiation (horizontal and in-plane) and ambient temperature comparisons with real measurements are shown in Figs. 2a and 2b and the statistical results are presented in Table 2. The method works well for horizontal irradiation, with a monthly RMSE of 2.8% and MBE of 1.5%. Ambient temperature is also well described with RMSE and MBE of 0.15 and −0.14% (in Kelvin), respectively.

A noticeable deterioration in the statistic metrics is noted for in-plane irradiation, which is primarily due to the sub-models used in the process of translation of global horizontal irradiation into the plane of array. This is to be anticipated, as separation

![Fig. 1](image1)

Fig. 1 Plots of
a) Training sets around the missing period taking as starting points the 31st of May and the 1st of July 2014, and going backwards and forwards in time, respectively
b) Fitting curve for the optimum training set (15 days backwards and 26 days forwards) for module A

![Fig. 2](image2)

Fig. 2 Comparison of
a) Global horizontal (GHI) and plane-of-array irradiation (POA)
b) Ambient temperature for measured and interpolated data throughout a year (2014)

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Statistical results for annual analysis with measured and interpolated climatic data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monthly</td>
</tr>
<tr>
<td>Ambient temperature, K</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.43</td>
</tr>
<tr>
<td>MAE</td>
<td>0.40</td>
</tr>
<tr>
<td>MBE</td>
<td>−0.40</td>
</tr>
<tr>
<td>RMSE, %</td>
<td>0.15</td>
</tr>
<tr>
<td>MAE, %</td>
<td>0.14</td>
</tr>
<tr>
<td>MBE, %</td>
<td>−0.14</td>
</tr>
</tbody>
</table>

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
(global irradiance to beam and diffuse) and translation algorithms add a high percentage random and bias error, which varies amongst different models and locations [26]. However, there is potential for improvement in determining the optimum separation model for the UK climate. Thus generally, the results depend strongly on the climatic profile of the location and the choice of models. In-plane irradiation is generally underestimated, with some days giving better results than others. A further analysis based on different irradiation bins and clearness indices shows that the random error derives mainly for low irradiation and partly cloudy days as seen in Figs. 3a and b. Clearness index was calculated using [27] and the days were classified according to Gul et al. [28].

The width and the number of the irradiation bins were adjusted considering the frequency of irradiation values, so that RMSE in different bins is affected by the same number of observations. Lower irradiation values present a higher percentage RMSE which however is small in terms of absolute energy yield (Wh). The data points in Fig. 3a represent the calculation bias, which changes according to clearness index. It seems that the method tends to underestimate higher irradiance (negative MBE) which present a higher bias error, whereas for irradiance values lower than 100 W/m² the result is slightly overestimated (positive MBE). This is due to the separation into beam and diffuse algorithms which tend to overestimate diffuse radiation for days with higher clearness index. Partly cloudy days contribute significantly to the overall error for the majority of the bins.

3.2 Inferring missing meteorological and electrical data

Concurrent energy yield readings and a climatic dataset were utilised containing a 1 month period of missing data (June 2014), during which neither of the above information was available. In order to validate the modelling results, this period was completely removed from the initial dataset and was treated as the ‘missing’ period. To calculate energy output, (3) is employed twice. Initially, it is used with hourly measured data of in-plane irradiation, module temperature and energy output to extract the model coefficients using a training period as defined using the training algorithm described in 0. Then, it is applied again to calculate the energy output for the missing month, using interpolated climatic data only for this period. Aggregated irradiation is calculated using hourly sums of irradiance. Module temperature for the missing period is calculated using (2) with interpolated irradiation and ambient temperature as input parameters. Comparisons between the obtained results and actual measurements for the missing period are shown in Figs. 4 and 5, followed by the statistical results in Tables 3 and 4.

The results for ambient temperature show that it can be interpolated to the location of interest with a very small MBE and RMSE. This is to be expected, as temperature is temporally and spatially more homogeneous than irradiance over the same distance for the UK climate. MBE increases for module temperature due to error propagation from both in-plane irradiation (inherent underestimation) and ambient temperature, but the effect

![Fig. 3 Statistical analysis on hourly irradiation bins](image)

**Fig. 3** Statistical analysis on hourly irradiation bins

*a* %RMSE and %MBE 
*b* Normalised contribution to RMSE of hours with different clearness index $K_t$

![Fig. 4 Comparison of daily results for the missing month (June 2014) for modelled and measured in-plane irradiation (POA) and average module temperature for](image)

**Fig. 4** Comparison of daily results for the missing month (June 2014) for modelled and measured in-plane irradiation (POA) and average module temperature for

*a* Module A
*b* Module B
Table 3  Statistical results for in-plane irradiation and module temperature comparisons

<table>
<thead>
<tr>
<th></th>
<th>Module A</th>
<th>Module B</th>
<th>Module A</th>
<th>Module B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daily</td>
<td>Monthly</td>
<td>Daily</td>
<td>Monthly</td>
</tr>
<tr>
<td>In-plane irradiation (kWh/m²)</td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Daily</td>
<td>0.70</td>
<td>8.3</td>
<td>2.78</td>
<td>2.24</td>
</tr>
<tr>
<td>Monthly</td>
<td>-0.28</td>
<td>-8.3</td>
<td>-0.76</td>
<td>-2.24</td>
</tr>
<tr>
<td>RMSE, %</td>
<td>14.9</td>
<td>5.9</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td>MAE, %</td>
<td>10.5</td>
<td>5.9</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>MBE, %</td>
<td>-5.9</td>
<td>-5.9</td>
<td>-0.76</td>
<td>-0.75</td>
</tr>
<tr>
<td>RMSE, %</td>
<td>14.9</td>
<td>5.9</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td>MAE, %</td>
<td>10.5</td>
<td>5.9</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>MBE, %</td>
<td>-5.9</td>
<td>-5.9</td>
<td>-0.76</td>
<td>-0.75</td>
</tr>
</tbody>
</table>

Table 4  Statistical results for energy output and PR for the two modules

<table>
<thead>
<tr>
<th></th>
<th>Module A</th>
<th>Module B</th>
<th>Module A</th>
<th>Module B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daily</td>
<td>Monthly</td>
<td>Daily</td>
<td>Monthly</td>
</tr>
<tr>
<td>Energy output, kWh</td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Daily</td>
<td>0.15</td>
<td>1.81</td>
<td>0.040</td>
<td>0.0002</td>
</tr>
<tr>
<td>Monthly</td>
<td>0.11</td>
<td>1.81</td>
<td>0.026</td>
<td>0.0002</td>
</tr>
<tr>
<td>RMSE, %</td>
<td>14.5</td>
<td>5.92</td>
<td>4.39</td>
<td>0.016</td>
</tr>
<tr>
<td>MAE, %</td>
<td>10.8</td>
<td>5.92</td>
<td>2.90</td>
<td>0.016</td>
</tr>
<tr>
<td>MBE, %</td>
<td>-5.92</td>
<td>-5.92</td>
<td>-0.97</td>
<td>-0.016</td>
</tr>
<tr>
<td>Performance ratio, PR</td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Daily</td>
<td>0.040</td>
<td>0.0002</td>
<td>0.042</td>
<td>0.007</td>
</tr>
<tr>
<td>Monthly</td>
<td>-0.009</td>
<td>-0.0002</td>
<td>-0.0024</td>
<td>-0.007</td>
</tr>
<tr>
<td>RMSE, %</td>
<td>4.39</td>
<td>0.016</td>
<td>4.85</td>
<td>0.86</td>
</tr>
<tr>
<td>MAE, %</td>
<td>2.90</td>
<td>0.016</td>
<td>3.44</td>
<td>0.86</td>
</tr>
<tr>
<td>MBE, %</td>
<td>-0.97</td>
<td>-0.016</td>
<td>-0.29</td>
<td>-0.96</td>
</tr>
</tbody>
</table>

Fig. 5  Comparison of daily modelled and measured energy output and PR for the missing month (June 2014) for

a  Module A
b  Module B

a  Module A
b  Module B

Fig. 6  Scatter diagrams for module A, of hourly modelled and measured for the missing month (June 2014) for

a  In-plane irradiation
b  Energy output
is very small with a maximum deviation of about 5° and average deviation of about 2°.

The results for energy output errors are a propagation of irradiation and module temperature. The negative bias which arises primarily due to irradiation is evident also in energy output and in terms of absolute RMSE for energy output, that is 1.8 and 1.5 (in kWh) for modules A and B, respectively. Crucially, the monthly PR can be predicted with a very small error for both cases and more specifically, with RMSE of about 0.0002 and 0.007 for modules A and B, respectively. A comparison of the above graphs, shows that for days where temperature is particularly low, measured PR is high and may slightly exceed 1.0, leading to a small deviation from the modelled value. Module temperature plays a significant role in modelling the energy output, which however is not directly evident in the PR. Namely, if in-plane irradiation is overset (Fig. 4), modelled energy output (Fig. 5) is affected by both in-plane irradiation and module temperature rise. The modelled result is very close to the measured energy output (where actual in-plane irradiation is lower) and thus, modelled PR is slightly lower than the measured value. This behaviour is particularly evident for days with lower average module temperature (i.e. low ambient temperature and/or windy days). In Figs. 6a and b, the scatter diagrams of modelled and predicted in-plane irradiation and energy output show that the discrepancy is low with a relatively small number of outliers in both cases being the main reason for the higher RMSE values for in-plane irradiation and energy output. This discrepancy is largely diminished with regards to the daily and monthly results for PR (see (1)).

4 Conclusions
A method to replenish missing meteorological and electrical data (not) obtained during PV system operation was developed and validated for a polycrystalline and a monocrystalline PV module. The method is based on interpolating meteorological data and translating it to the local climate (namely in-plane irradiation and module temperature) governing the device performance. The local climate data are then used to calculate the electrical output of the system. This approach is validated against data from a precision measurement system.

There are noticeable differences in terms of the absolute energy production, while the estimation of the PR shows excellent agreement for both PV technologies. This means that the key property for assessing system quality can be replenished accurately with the given method and thus this is sufficient for evaluating real-time continuous performance. The PR is the key parameter required for warranty verification, and the method is highly successful for achieving this. Detailed investigation of the relative underestimation of the energy yield of a system identified that the error is almost exclusively due to the irradiance translation to plane of array and thus further efforts will focus on this. The largest bias is seen for high irradiance conditions, while the highest scatter is seen for lower irradiances but the reasons for this are not yet entirely clear.

Finally, the next step is to scale this study up to a specific set of data, such as a number of PV systems in close geographic proximity, for example, over a given urban area. This will require simulation of the entire PV system, i.e. take into consideration inverter and shading effects. This task is currently ongoing [29] and together with the work carried out in this study, it will enable automated analysis of domestic systems over a larger area.

5 Acknowledgments
This work has been conducted as part of the research project ‘PV2025 - Potential Costs and Benefits of Photovoltaic for UK Infrastructure and Society’ project which is funded by the RCUK’s Energy Programme (contract no: EP/K02227X/1).

6 References
6 ‘Met Office Integrated Data Archive System (MIDAS) Land and Sea Surface Stations Data (1853-CURRENT)’, http://catalogue.ceda.ac.uk/uuid/220a656152148d5c1cc075f8a32324d0, last accessed 3 June 2015