ESTIMATION OF THE DEGRADATION RATE OF FIELDED PHOTOVOLTAIC ARRAYS IN THE PRESENCE OF MEASUREMENT OUTAGES

Alexander Phinikarides, George Makrides and George E. Georgiou
Photovoltaic Technology Laboratory, FOSS Research Centre for Sustainable Energy, Department of Electrical and Computer Engineering, University of Cyprus.
1. Panepistimiou Avenue, 2109 Aglantzia, Nicosia, Cyprus.
(Telephone: +357 22 894398, Fax: +357 22 895370, email: phinikarides.alexander@ucy.ac.cy)

ABSTRACT: This paper presents a study of the sensitivity of the estimated energy degradation rates of grid-connected crystalline Silicon (c-Si) PV arrays to measurement outages and missing data. The arrays under study have been operating side-by-side since June 2006 in Nicosia, Cyprus. Fifteen-minute average field measurements over their first ten years of operation were used to create a data set, in which invalid measurements were removed and logged system downtimes were corrected based on past performance. The resulting data set was randomly sampled to select data points which were designated as Missing Completely at Random (MCAR) in order to create unbiased artificial outage periods. The same data set was treated with missing data imputation techniques such as imputation by the mean, linear interpolation and imputation by bootstrap. The resulting data sets were then analysed with linear regression (LR), classical seasonal decomposition (CSD) and regARIMA models to extract the trend and estimate the energy degradation rates. The results have shown that regARIMA was the most robust method for up to 10% of missing data, the LR and CSD methods were sensitive starting from 2% of missing data and that imputation by the bootstrap enhanced the accuracy of the estimated degradation rates up to 20% for regARIMA and 40% for LR.

Keywords: degradation, photovoltaic, performance, monitoring, outages, imputation.

1 INTRODUCTION

A major requirement for the uptake of PV is assurance of the lifetime energy yield and formulation of strict warranties in order to reduce investment risk. This requires accurate evaluation of the energy degradation rate, $R_{de}$, of fielded PV systems [1]. The term energy degradation rate includes shading and soiling effects, module mismatches and module degradation and is a positive quantity. Factors such as length of field exposure, module technology, measurement accuracy and precision and measurement outages all affect the degradation rate estimates [2].

Measurement outages cause missing measurements and incomplete data sets. The outages can be due to data logger faults, measurement noise, data transmission errors, connection errors and data storage faults. Missing data is also caused by discarding invalid measurements and outliers due to inverter clipping, misconfigured sensors, sensor drifts and invalid calibration, array or irradiance sensor shading and soiling and PV system downtimes. Additionally, outliers can be determined from abrupt changes and high-frequency content in the measurements and also from highly volatile measurements (higher than the response time of the sensors.)

As the energy degradation rate is estimated with statistical methods, a multitude of statistical analyses were developed to test the sensitivity and the bias introduced to the resulting $R_{de}$ estimated from data sets with missing measurements. In the presence of measurement outages, these incomplete data sets can bias the $R_{de}$ estimate, either by exacerbating the effect of the outage factors when the missing data point is of an instance of optimal operation, or by overestimating PV system performance and therefore understimating degradation, when the missing data point is of a period of increased system losses.

2 METHODOLOGY

2.1 Experimental setup

At the outdoor PV Technology test site of the University of Cyprus (UCY), grid-connected PV systems of different technologies and approximately 1 kWp capacity each, were installed and commissioned in June 2006. The performance of each PV system and the prevailing meteorological conditions were recorded according to IEC 61724 [3] at a resolution of one second and stored as fifteen-minute averages. The meteorological measurements include the global irradiance on the plane of array (POA), $G_i$, wind direction, $w$, and wind speed, $Sw$, as well as ambient, $T_{amb}$, and module temperature, $T_m$. The electrical measurements include maximum power point (MPP) DC current of the array, $I_a$, voltage, $V_a$, and power, $P_{va}$, and AC power to the utility grid, $P_{TU}$.

2.2 Data sets of observed measurements

Measurements of the DC array power at MPP, $P_a$, of three different c-Si PV systems and global irradiance on the POA, $G_i$, from a Kipp & Zonen CM21 pyranometer were selected for the first ten years of operation (June 2006 – May 2016). The long testing period ensured that the estimated energy degradation rates represented their long-term operation and were not skewed by reversible degradation factors [4].

Invalid measurements and outliers were removed from the $P_a$ and $G_i$ time series by preset thresholds and maintenance logs were used to correct periods of system and sensor downtimes. The downtimes were corrected using past measurements for short outage periods (less than one day). For longer periods, the outages were not corrected to avoid introducing bias. Calibration factors were also applied to the $P_a$ and $G_i$ measurements. The resulting data sets were defined as the complete data sets of observed values, $D^{obs}$.

The complete data sets were used to create time series of monthly Performance Ratio, $PR$ [3], as shown in Figure 1. The $PR$ plots show the typical seasonal profile
of c-Si PV systems, with high PR in the winter and lower in the summer, due to higher module temperatures.

![Figure 1: Monthly Performance Ratio (PR) of the three c-Si PV arrays under study from June 2006 to May 2016](image)

2.3 Generation of artificial missing data

A Monte Carlo approach was employed, in order to create data sets with artificial missing data from the complete data set of observed values, $D^{obs}$, for each PV system. The index of the complete data set was randomly sampled at different levels in order to select artificial outage periods. Random sampling was performed from 1 to 40% of the total amount of data in the complete data set. The data points, $M$, were then designated as missing from the complete data sets. The resulting incomplete data sets contained instances where all variates were missing at once (case deletion), to simulate multiple failures in the whole measurement chain.

Since random sampling was utilized, an unbiased set of data points was selected. This method represents the case where data is Missing Completely at Random (MCAR) [5] and the distribution of missing data points did not depend on either the observed values or the missing values (see Eq. 1.)

$$P(M \mid D^{obs}) = P(M) \quad (1)$$

Lastly, as with the case of the complete data, the incomplete data sets were used to create monthly PR time series for each percentage of missing data. In total, forty incomplete data sets were created for each of the PV systems in Figure 1.

2.4 Imputation of missing data

As the estimation of the degradation rate relies on statistical analysis, the statistical properties of the time series must be retained without introducing bias.

When using PV models such as PVUSA [6], single-point efficiency or others [7] to interpolate missing values, all the required explanatory variables may not always be available and may not always be measured (irradiance, module temperature, wind speed, humidity.) Additionally, since data logger, data transmission and storage related errors affect all measured variables, PV models were not considered in this study. Instead, the handling of missing data was based on univariate imputation of missing $G_t$ and $P_A$ data points, where the basic assumptions were that the variables were normally distributed and that the pattern of missing data was independent of the underlying data set.

The missing data points were imputed with a) the unconditional mean of the variable, b) linear interpolation and, c) bootstrapping in order to fill in the gaps in the fifteen-minute data sets.

Firstly, the missing data points were imputed by the unconditional mean of the data set. This ensured that the probability distribution of the measurement variable remained the same, but at the expense of large errors due to the seasonal profile of $P_A$ and $G_t$.

Similarly, the same missing data points were imputed by linear interpolation with time, in order to better model the effect of the trend present in the complete data set (see Eq. 2.)

$$P_A = \beta_1 + \beta_2 t + \varepsilon \quad (2)$$

where $\beta_1$ and $\beta_2$ are regression coefficients, $t$ is the time and $\varepsilon$ are the residuals.

2.5 Imputation by the bootstrap

In addition to imputation by the mean and linear interpolation, the bootstrap method was employed in order to use as much information from the probability distribution of each variable and be able to assess the uncertainty of the imputed values [8]. The bootstrap method relies on sampling from the posterior distribution and replacing missing data points with the sampled values. The algorithm then re-evaluates and resamples the posterior distribution, and replaces the missing data points with the newly sampled values. This iterative procedure is performed until the Expectation Maximization algorithm [9] converges. The result is that the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data, for each data point.

The energy degradation rate of PV arrays was observed to vary smoothly over continuous exposure time [10], either linearly or with inflection points. In order to incorporate linear degradation in the bootstrap algorithm, the effect of outdoor exposure time was added by building a general model of variables across time by creating a sequence of polynomials of the time index [8]. In this case, observed values close in time to any missing data had the largest contribution to the imputation of the missing data point. The $P_A$ data sets for each PV system were therefore described as a first-order polynomial of time (see Eq. 3):

$$P_A \propto \beta_1 t \quad (3)$$

The algorithm was thus modified so that the imputed data sets, constructed from each run of the bootstrap algorithm, were regressed on Eq. 3 and the missing data points interpolated based on this linear relationship.

2.6 Estimation of the energy degradation rate

A large number of data sets were created by applying the methodology described in the previous sections. For each PV system, 161 different data sets were analysed, in order to estimate the energy degradation rate: a) one complete data set, b) forty incomplete data sets with 1-40% of missing data, c) forty data sets imputed by the
mean, d) forty data sets imputed by linear interpolation and, e) forty data sets imputed by the bootstrap.

Each data set of $G_t$ and $P_a$ values was used to create monthly $PR$ time series which were then analysed with different statistical methods over the first ten years of field exposure [11], [12]. The $PR$ time series of each PV array was seasonally adjusted with Regression models with ARIMA errors (regARIMA) [13], [14] and classical seasonal decomposition (CSD) in order to separate it into the trend, the seasonal and the irregular components and was also modelled with linear regression (LR) using ordinary least squares.

CSD was used to fit an additive model on the data and separate the trend, $T$, the seasonal, $S$, and the residual, $R$, components (see Eq. 4):

$$ Y_t = T_t + S_t + R_t $$  (4)

RegARIMA models were applied on the time series through the X-12-ARIMA algorithm. These model the mean of the time series by a linear combination of regressors and the residual component by an ARIMA process (see Eq. 5):

$$ Y_t = \sum_{i} \beta_i x_{it} + z_t $$  (5)

where the residual component, $z_t$, is modelled by the seasonal ARIMA model given in Eq. 6:

$$ \Phi(T) \Phi_S (T_S) \nabla^d \nabla_S^D z_t = \theta(T) \theta_S (T_S) e_t $$  (6)

Lastly, the completed data sets were used to estimate the linear energy degradation rate, $R_{DE}$, using linear regression (see Eq. 7):

$$ R_{DE} = -\alpha * 12 * 100\% $$  (7)

where $\alpha$ is the slope of the linear equation. The monthly $R_{DE}$ was multiplied by 12 to convert it to an annual rate.

3 RESULTS

The results of the analysis for 1–40 % of sampled data points missing completely at random are shown in Fig. 2, Fig. 3 and Fig. 4 for the LR, CSD and regARIMA methods respectively. It can be seen that imputation by the mean and linear interpolation performed very poorly for all PV systems as these two methods underestimated the $R_{DE}$ consistently with increasing amount of missing data. On the other hand, imputation by the bootstrap has been shown to provide a large improvement to all three degradation estimation methods. Therefore, the bootstrap method provided robustness to the estimation, regardless of the degradation estimation method employed. For LR, the bootstrap provided robustness for up to 20 % of missing data for the ucy07 and ucy11 systems, whereas for the ucy08 system, the estimates were stable even at 40 % of missing data.

The results of applying CSD were shown to be very sensitive to the amount of missing data and whether imputation was used or not. Without using imputation, the estimates were underestimated by 0.22 %/year for the ucy07 system and overestimated by 0.05 %/year for the ucy08 and ucy11 systems. By using the bootstrap, the under/overestimation was reduced to 0.05 %/year for all systems.

In the case of regARIMA, the results were shown to be the most robust to missing data for all systems, regardless of whether imputation was employed. For the ucy07 and ucy11 systems, the maximum differences without imputation were 0.05 %/year and for the ucy08 the maximum difference was 0.02 %/year. When imputed by the bootstrap, the differences were slightly reduced.
4 CONCLUSIONS

The results of the investigation into the energy degradation estimation’s sensitivity to missing data have shown that missing data as defined in this work greatly affected the estimation of degradation rates.

To mitigate the effects of missing data, imputation by the bootstrap was shown to be the most successful in providing robust energy degradation rate estimates. In contrast, imputation by the mean or linear interpolation were incapable of providing any improvement and resulted in worse estimates than not applying any imputation.

Finally, the application of statistical methods to estimate the degradation rate was also shown to be sensitive to the amount of missing data. regARIMA was the most robust method, whether imputation was applied or not. LR was more robust when missing values were imputed by the bootstrap and CSD was the most sensitive method to missing data.

5 ACKNOWLEDGEMENTS

This work was funded through the IPERMON project (KOINA/SOLAR-ERA.NET/1214/08) which was co-financed by the European Regional Development Fund and the Republic of Cyprus through the Cyprus Research Promotion Foundation (DESMI 2009-2010)

6 REFERENCES