A method for detecting malfunctions in PV solar panels based on electricity production monitoring

Fermín Mallor¹, Teresa León², Liesje De Boeck^{3,4}, Stefan Van Gulck⁵, Michel Meulders^{6,7}, Bart Van der Meerssche^{7,8}

¹Public University of Navarre (UPNA), Statistics and Operations Research Department, Campus Arrosadia, 31006 Pamplona (Navarre), Spain, mallor@unavarra.es

²University of Valencia, Department of Statistics and Operations Research, C/ Doctor Moliner 50, 46100 Burjassot, Spain, teresa.leon@uv.es

³KU Leuven Campus Brussels, Faculty of Economics and Business, Research Centre for Operations Management, Warmoesberg 26, 1000 Brussels, Belgium, liesje.deboeck@kuleuven.be

⁴KU Leuven, Faculty of Economics and Business, Research Centre for Operations Management, Naamsestraat 69, 3000 Leuven, Belgium,

⁵KU Leuven Campus Brussels, Faculty of Economics and Business, Research Centre for Mathematics, Education, Econometrics and Statistics, Warmoesberg 26, 1000 Brussels, Belgium, stefan.vangulck@kuleuven.be

⁶KU Leuven Campus Brussels, Faculty of Economics and Business, Research Centre for Mathematics, Education, Econometrics and Statistics, Warmoesberg 26, 1000 Brussels, Belgium, michel.meulders@kuleuven.be

⁷KU Leuven, Faculty of Psychology and Educational Sciences, Research Group on Quantitative Psychology and Individual Differences, Tiensestraat 102, 3000 Leuven, Belgium

⁸FLUKSO bvba, 3150 Wespelaar, Belgium, bart.vandermeerssche@flukso.net

Abstract

In this paper a new method is developed for automatically detecting outliers or faults in the solar energy production of identical sets (sister arrays) of photovoltaic (PV) solar panels. The method involves a two-stage unsupervised approach. In the first stage, "in control" energy production data are created by using outlier detection methods and functional principal component analysis in order to remove global and local outliers from the data set. In the second stage, control charts for the "in control" data are constructed using both a parametric method and three non-parametric methods. The control charts can be used to detect outliers or faults in the production data in real-time or at the end of the day. As an illustration, the method is applied to analysis of the real energy production data of six sets of "identical" PV solar panels over a period of three years. Tests indicate that the proposed method is able to successfully detect a reduction in efficiency in one of the solar panel sets by up to 5%. Control charts based on parametric and non-parametric methods both show good performance results.

Keywords

Keywords: anomaly detection; photovoltaic solar panel sister arrays; electricity production monitoring; statistical quality control

1. Introduction

In recent decades, renewable energy sources have gradually risen to prominence for different reasons, such as rising energy prices, an increase in greenhouse gas emissions, depletion of fossil fuels, etc. (De Boeck et al., 2016). Political agreements made at an international level encouraged this evolution and originated from the Kyoto protocol. The European Union has always played a prominent role in this process. In 2007 it reached an agreement to increase its share of renewable energy in total production to 20 % by 2020 (European Commission, 2007). Photovoltaic (PV) solar panels were an essential component in reaching this target. Ever since, governments worldwide have established indirect and direct support systems to promote renewable energy sources as an alternative to traditional energy sources.

Parallel with the evolution in the use of renewable energy sources, we also observe a trend of increased use of energy measuring devices. This is not surprising given the increased awareness of energy use and its related costs among energy consumers that goes along with evolution to a more ecological and sustainable world. Indeed, modern sensors already allow online monitoring of e.g., energy production of PV panels. However, literature related to outlier detection tools in this energy production have only come into the picture during the last half decade. Nonetheless, outlier detection in general is not new (see e.g., Aguinis et al., 2013). Indeed, as outlier detection can potentially save a lot of losses (in terms of money, time, production, lives, etc.), it has been an important research area for a long time. Popular application areas include credit card fraud detection, computer intrusion detection and medical applications.

In this paper, we specifically address the problem of monitoring the solar energy production of identical sets (i.e., sister arrays) of PV panels (i.e., sets sharing the same characteristics such as location, orientation, inclination slope, age, module material, BOS components, local thermal environment, etc.). A PV array is here defined as a parallel set of PV strings, a string being an assembled set of PV modules or panels; the different strings in an array share the same voltage and each array is connected to one inverter (Zhao et al., 2013). In what follows we will refer to arrays as “sets”. Note that as differences in the sets will affect the effectiveness of our method, the focus in our paper is on systems having identical arrays (such as PV farms). The aim is then to develop a method of detecting all malfunctions of one of the PV system elements (that affect the generated number of kWh in PV systems) as soon as possible in order to facilitate a timely response. In this way, the economic impact (in terms of costs related to this malfunctioning) can be minimized and the reliability of PV panels as a renewable system can be guaranteed. To this end, we will develop a two-stage unsupervised method that starts from historical PV panel energy production data. In the first stage, we build functional observations and aim at the following: normalized energy production over all sets of panels and an equal number of observations for all sets and all days, detecting outlying observations and failures (in the recording/measurement devices, in the functioning of panels and/or invertors) as well as seasonal shadowing phenomena. This first stage procedure leads to a representative sample of days with stable energy production, i.e. data representing the “in control” behavior of the energy system. In the second stage, we rely on statistical quality control methods to

build control charts using the “in control” observations. To this end, we develop one parametric and three additional non-parametric methods. The control charts are able to monitor the energy production and detect outliers in real time or at the end of the day. We will apply the two-stage method to a real-life case. Data for this case were provided by Flukso.

An important potential application area of outlier detection is abnormal energy use in general and PV panels functioning more specifically. Some papers focus on energy use in buildings (see e.g., Dodier et al., 1999; Xiaoli et al., 2010; Capozzoli et al. 2015). Because there exists a trend of increasing energy demand in buildings and long-term use of buildings, the energy efficiency of these buildings as well as energy management is of prime concern for anyone wanting to identify energy savings. Other papers discuss energy yield monitoring and fault detection in PV panel functioning. Firth et al. (2010) demonstrated that faults can potentially cause significant energy losses in domestic PV systems. A review of the three major failures in PV arrays (ground faults, line-to-line faults, and arc faults), their effect on the PV array functioning as well as limitations to existing detection and mitigation technologies can be found in Alam et al. (2015). A comprehensive review of existing fault detection and classification methods in the DC side of PV strings/arrays can be found in Akram and Lotfifard (2015). Most of the models used for outlier detection in PV panel functioning compare monitored data from the PV system with model prediction results in order to identify faults when significant differences are observed between both data sets (Silvestre et al., 2013). These models typically develop thresholds below or above which fault signals arise (Akram and Lotfifard 2015). Different techniques such as parametric models, artificial intelligence techniques or commercial simulation packages can be used to develop the prediction models (Platon et al., 2015). Artificial intelligence techniques (e.g., neural network, and expert systems) are frequently used within the machine-learning field. In PV panel fault detection, machine-learning algorithms learn about the relation between input and output parameters of the PV system and then the trained models are used to detect and classify faults (Akram and Lotfifard, 2015). Model training and prediction can be based on data obtained from remote monitoring (see e.g., Stettler et al., 2006; Drews et al., 2007, Chouder and Silvestre, 2010, Gokmen et al., 2012), e.g., on meteorological and satellite observation data (Tadj et al., 2014), other meteorological data bases (Quesada et al., 2011) or historical data (Platon et al. 2015). Alternatively, forecasts can also be based on actual monitored data, i.e., real time data (see e.g., Chouder and Silvestre, 2010; Chouder et al., 2013; Sylvestre et al., 2013). Also what is measured can differ and mostly depends on the fault that one wants to detect. Fault classification and detection can rely on e.g., the rate of power or energy losses in the PV system (see e.g., Drews et al., 2007; Chouder and Silvestre, 2010; Silvestre et al., 2013; Silvestre et al., 2014; Silvestre et al., 2015), heat exchange and module temperature (see e.g., Vergura et al., 2012; Hu et al., 2013), energy output of inverters (see e.g., Vergura et al., 2009), ratio between DC and AC power (see e.g., Chine et al., 2014), etc. There are also methods that use external devices for fault detection such as signal generators, and LCR [inductance (L), capacitance (C), and resistance (R)] meters (Akram and Lotfifard, 2015). Finally, there are methods (see e.g., Zhao et al., 2013) that rely on differences between measurements from adjacent strings for fault detection in PV strings (Akram and Lotfifard, 2015). So the paper that bears most comparison with ours is the paper by Zhao et al. (2013). They also do not use weather data, information from the manufacturer of the

solar panels about the physical characteristics, etc. in order to perform their analysis but they use sister “components” (i.e., sister strings). They compare normalized currents in the different strings that are simultaneously measured to detect a fault in one of the strings. To this end they calculate the average, standard deviation, and the three quartiles based on a dataset with three or eight values, which are the number of strings in their small experimental setup. They study the following outlier detection rules: 3-sigma rule (classify an observation as an outlier if it lies more than three standard deviations from the mean), Hampel identifier (classify an observation as an outlier if it lies further away from the median than three medians of the absolute deviations around the median, Boxplot rule (classify an observation as an outlier if $x > Q3 + 1.5IQR$ or $x < Q1 - 1.5IQR$, see Table 2 for the meaning of $Q1$, $Q3$ and IQR). A clear difference with our approach is that we use historic production data to derive the “in control” reference dataset of the production range of arrays (i.e., the difference between the array with largest production output and smallest production output at a time point). Next we propose to apply control charts (e.g., 3-sigma rule) to the production range of new observations in order to determine whether the observation is an outlier (and hence to infer whether a fault has occurred). In contrast Zhao et al. (2013) do not derive “in control” data, but they directly apply the outlier detection method to data from a specific time point. We believe that this “in control” reference dataset is an important part of the methodology to get reliable error detection. So our method detects outliers in the range of current production which, once the control chart is calibrated, does not need any extra information. Other methods based on detecting outliers (understood as large deviations in the expected production) need some reference pattern. This pattern (see supra) could be obtained by simulating the expected production from weather forecasts or from some type of online weather monitoring, or from historical records of both production and weather conditions. The main advantage of our method is thus simplicity.

Our paper is organized as follows. Section 2 will zoom in on the methodology. In the third section we apply the two-stage method to real-life data, followed by a discussion of the results. Section 4 will end with the most important conclusions of the paper and some points for future research.

2. Methodology

We start our methodology section by explaining our way of working, defining the notation we will use throughout this section and by providing an overview of the methodology stages in Section 2.1. Section 2.2 zooms in on the first main stage of the methodology where we aim at obtaining a representative sample that describes the energy production process in a stable state. In Section 2.3 we develop control charts to monitor energy production, based on the representative sample of in control state production obtained from the first stage. The nomenclature we will use in this and the following sections is represented in Table 1.

Symbol	Definition
AC	alternating current
$C(d)$	T vector of cumulative production range at day d
$C(t, d)$	cumulative production range at time interval t at day d
d	day
DC	direct current

EV	extreme value
FDA	functional data analysis
FPCA	functional principal components analysis
FPC	functional principal component
GEV	generalized extreme value
IQR	interquartile range (Q3-Q1)
K	index of a PV panel
K	number of PV panels
LCR	inductance, capacitance and reductance
$L(d)$	L2-norm of $C(d)$
$m(d)$	pointwise mean of $\{C(d)\}_{d \in N(d)}$
$max(d)$	pointwise maximum of $\{C(d)\}_{d \in N(d)}$
$min(d)$	pointwise minimum of $\{C(d)\}_{d \in N(d)}$
N	number of days
$n(d)$	number of days in $N(d)$
$N_s(d)$	neighborhood of days (from $d-s$ till $d+s$) with cumulative range vectors similar to $C(d)$
PC	principal component
PCA	principal components analysis
PV	photovoltaic
$Q1$	first quartile
$Q3$	third quartile
$Q1(d)$	pointwise first quartile of $\{C(d)\}_{d \in N(d)}$
$Q3(d)$	pointwise third quartile of $\{C(d)\}_{d \in N(d)}$
$R(d)$	T vector of production range at day d
$R(t, d)$	production range (maximum production minus minimum production) in a set of K panels at a time interval t at day d
s	time window of neighborhood $N(d)$
$sd(d)$	pointwise standard deviation of $\{C(d)\}_{d \in N(d)}$
T	a time interval
T	number of time intervals
UCL	upper control limit
$UCL(d)$	upper control limit vector for the cumulative ranges of days in $N(d)$
$UCL(t, d)$	component of $UCL(d)$ corresponding to time interval t
$ULGEV$	upper control limit derived from GEV distribution
$X_k(t, d)$	production (kWh) of panel k during the time interval t at day d
Y	a year
Y	number of years
$\alpha_1, \alpha_2, \alpha_3$	quantiles determining the probability of a Type-I error
μ	population location parameter of GEV distribution
$\hat{\mu}$	estimate of location parameter of GEV distribution
ξ	population shape parameter of GEV distribution
$\hat{\xi}$	estimate of shape parameter of GEV distribution
σ	population scale parameter of GEV distribution
$\hat{\sigma}$	estimate of scale parameter of GEV distribution
$GEV(\mu, \sigma, \xi)$	Generalized Extreme Value distribution with location parameter μ , scale parameter σ and shape parameter ξ

Table 1. Nomenclature.

2.1 Explanation, notation and overview

PV panels' energy production is a cyclical phenomenon over a one-day time span. Historical energy production data make it possible to develop a daily production profile for

each of the panel sets whose production is monitored. We assume availability of the following historical data: the generated number of kWh (i.e., the P_{DC} value) for each m minute interval of each set of PV panels for a sample of days. More specifically, we have available the production data $X_k(t, d)$ of T time intervals per day of K solar panel sets during N days represented in Eq. (1) as:

$$\{X_k(t, d); t = 1, \dots, T; k = 1, \dots, K\}_{d=1}^N, \quad (1)$$

with t the observed time interval, T the number of time intervals per day, d the observed day, N the number of monitored days, k the solar panel set and K the total number of solar panels sets. In what follows, we assume m to take on the value 5 (minutes) without loss of generality. When sets of PV panels are identical, energy production for each set is expected to be similar. Differences should be small and only relate to random causes of variation (e.g., minor differences in the material composition of the PV panels, atmospheric perturbations, etc.). When malfunctioning in a set of panels occurs, its production will differ from the other sets, leading to a significant increase in the range of production values coming from the whole set of PV panels. That is the reason why we develop an unsupervised statistical method to monitor the variable range (maximum production in the whole set minus minimum production in the whole set for each time interval t) in order to detect significant variations in this range. The range vector $R(d)$ for a given day d , $d = 1, \dots, N$, is defined in Eq. (2) as:

$$R(d) = (R(1, d), R(2, d), \dots, R(T, d)) \quad (2)$$

where

$$R(t, d) = \max_k \{X_k(t, d)\} - \min_k \{X_k(t, d)\} \quad t = 1, \dots, T. \quad (3)$$

Large values in this range will thus be indicative of large production deviations among the different sets of PV panels and thus of the malfunctioning of at least one of them. In fact, the detection of malfunctioning of some panel sets will be done by monitoring the cumulative range vector $C(d)$ because it is much smoother than the range function (see Section 3). For a given day d , $d = 1, \dots, N$, $C(d)$ is defined in Eq. (4) as:

$$C(d) = (C(1, d), C(2, d), \dots, C(T, d)) \quad (4)$$

where

$$C(t, d) = \sum_{i=1}^t R(i, d) \quad t = 1, \dots, T. \quad (5)$$

Note that both $R(d)$ and $C(d)$ can be modeled as functional random variables.

After having constructed the range and the cumulative range vectors, we distinguish two stages in the methodology (see Figure 1).

The first stage consists of obtaining a representative sample that describes the energy production process in a stable state, i.e., under control in terms of quality control terminology. To arrive at this representative sample, we use standard statistical tools and functional data analysis (FDA) to detect abnormal behavior. This behavior can be caused by outlying observations or failures (in the recording/measurement devices, in the

functioning of panels and/or inverters). Firstly, the boxplot of the L2-norm values of the cumulative ranges facilitates the detection of global malfunctioning. Secondly, the use of functional principal component analysis (FPCA) makes it possible to detect local temporal deviations in the production caused by e.g., shadowing phenomena.

In the second stage we develop a statistical method to monitor the energy production of a new day, based on the representative sample of the in control state production obtained from the first stage. For every day d , we define a neighborhood composed of days with cumulative ranges expected to be similar to $C(d)$. Then we develop upper control limits (UCLs) for the cumulative range which will be used to classify new observations as being generated by the “in control” or “out-of-control” production process. To this end, we combine tools from generalized regression methods and some ideas from statistical quality control. More specifically, we propose a regression model that uses extreme value distributions to estimate the upper limit of the cumulative range distribution in the “in control” state and three non-parametric methods.

Figure 1. Overview of the proposed methodology for PV solar panels malfunctioning for a set of solar panels.

2.2 First stage: selection of a representative sample of the “in control state” production process

We start our first stage with some pre-processing of the data. Days in the dataset with a very small number of measurements (e.g., fewer than six hours of observations) are removed because they usually indicate a malfunctioning of the recording device. In the event that certain sets of PV panels have a different capacity, we also rescale the energy production data of the sets to bring the energy production at the same level for all sets (see Section 3.1 for an example).

After the pre-processing of the data, the daily energy production $X_k(t, d)$ (see Eq. (1)) is stored in vectors with the same length (i.e., a length of T , which equals 288 when the energy production is recorded in five minute intervals). For intervals with no recorded production (e.g., night periods), a value of zero is assumed.

Thereafter, we construct the range vector (see Eq. (2)) and the cumulative range vector (see Eq. (4)).

2.2.1 Detection of global outliers using L2-norm and boxplots

In order to remove global outliers from the resulting dataset, the L2-norm, $L(d)$, is calculated for each vector $C(d)$ as represented in Eq. (6):

$$L(d) = \sqrt{\sum_{t=1}^T (C(t, d))^2}. \quad (6)$$

It is worth noticing that L2-norm is appropriate for comparing curves with similar shapes (Cerioli et al., 2005).

The boxplots of the $L(d)$ values can be visually inspected to detect global outliers. Common criteria of outlier detection (i.e., a data point below $Q_1 - 1.5IQR$ or above $Q_3 + 1.5IQR$) can be used to define outliers (where Q_1 , Q_3 and IQR denote the first quartile, the third quartile and the interquartile range respectively). Outlying days are analyzed and then removed from the dataset.

2.2.2 Detection of local outliers using Functional Data Analysis

The cyclical nature of the PV sets’ energy production data as well as the high frequency with which this production is monitored, favor the use of FDA to detect outliers. Although in practice each range vector consists of T discrete values, the vector can be turned into a continuous one. This can be performed in two ways: by interpolation or by smoothing. We opt for the smoothing method. A commonly used smoothing procedure consists of representing the discrete function through a linear combination of M known basis functions. Smoothed functional data are then created by fitting a B-spline basis to the set of range vectors $\{R(d)\}_{d=1}^N$. More precisely, we suggest using an order four B-spline basis function expansion with knots at sampling points. B-spline basis functions are polynomial segments joined end-to-end at certain argument values called knots, breaks or joint points. The segments have a specifiable smoothness across these knots. B-spline basis functions

have the advantage of being computationally inexpensive and offer a great deal of flexibility. To decide upon the number of basis functions M , a selection based on a cross-validation method is advisable. We have used the Generalized-Cross-Validation criterion implemented in the R-package “fda”. For more details on FDA, we refer to Ramsay and Silverman (2005) and Ramsay et al. (2014).

Thereafter, FPCA is applied to determine the main features in the variability of these smoothed range data (Ramsay and Silverman, 2005). The FPCA associated to a functional dataset is defined with the same properties as in the multivariate case. In such context, PCA finds the directions in the observation space in which the data have the highest variability. The central concept is to reduce the dimensionality of a dataset by transforming it into a new set of variables. The new variables are the principal components, which are uncorrelated and are ordered so that the first few components retain most of the variation present in all original variables. The functional case is analogous, each functional principal component (FPC) is specified by a principal component weight function $\xi(t)$ defined over range of the data and is uncorrelated to the others. The score z_d of a given element of the data set $R(t, d)$ can be obtained as $\int \xi(t)R(t, d)dt$. We obtain the FPCs using the R-Package fda (Ramsay et al., 2014). The visual inspection and interpretation of the graphic representation of the first FPCs helps to detect at which times the ranges differ greatly. Phenomena such as panels shadowing each other can be detected in this way. When the variability explained by the first r FPCs is close to 100%, the employment of a basis whose elements are these first r FPCs results in a useful finite-dimensional representation of the curves in which each curve is represented by its corresponding r -dimensional vector of scores. These scores can be used to detect local malfunctioning of the PV panel sets through their analysis using boxplots. The boxplots will provide outliers denoting extreme behavior in the variability pattern described by the corresponding FPC, pointing to possible malfunctioning of (at least one of) the panel sets or any issue that would require our attention.

2.3 Second stage: monitoring the PV energy production of a new day d

The first stage results in what is called “in state control” data in the field of quality control. These data are now used to define control charts in the wide sense of the term. The purpose of the control charts in this paper is to monitor the energy production and to detect malfunctions in the system of one or more PV panel sets. The monitoring is carried out by relying on the daily cumulative range vector $C(d)$. When $C(d)$ falls outside the control limits, it might be a signal of malfunctioning.

First note that the set of (random) vectors $C(d)$ within a year are not equally distributed because of the unique radiation pattern of each day in each year. Indeed, $C(\cdot)$ has different probability distributions on different days d in a specific year. However, observations of $C(d)$ at the same location in different years might be considered as a random sample of $C(d)$. Nonetheless, because the number of observations for $C(d)$ is often very small (i.e., less than three) for a given day d we define a set called the neighborhood of $C(d)$ with variables “similarly distributed” to the variable $C(d)$. One simple way of defining this neighborhood is by using a criterion of temporal proximity as indicated in Eq. (7) where s could be fixed at seven if a two-week stime window is used.

$$N_s(d) = \{C(w)\}_{w=d-s}^{d+s} \quad (7)$$

with $d=1,2,\dots,365$ (and where 1 corresponds to January 1st and so on).

When the PV energy production has been monitored for several years $y=1, 2, \dots, Y$, all the data recorded in different years are included in the neighborhood. If we denote by $n(d)$ the number of observations in $N_s(d)$, we have that $n(d) = (2s+1)*Y$. All observations of the variables in $N_s(d)$ are thus considered as a sample of the variable $C(d)$. The observations in a given day's neighborhood allow us to develop the control charts for the production of that day.

The control charts we will develop do not aim at detecting a situation where none of the PV panels sets is functioning. Our goal is to detect outliers or abnormal behavior of working PV panel sets. That is why we implicitly assume that at least one set of PV panels is working. As such, we will only focus on defining upper control limits. To this end, we will rely on one parametric method (see Section 2.3.1) and on three non-parametric methods (see Section 2.3.2).

2.3.1 Control chart based on the generalized extreme value distribution

Extreme value (EV) theory seeks to model the probability of events that are more extreme than any previously observed events. The probability density function for the GEV distribution with location parameter μ , scale parameter σ , and shape parameter $\xi \neq 0$ is represented in Eq.(8) as:

$$f(x|\mu, \sigma, \xi) = \frac{1}{\sigma} \exp \left\{ - \left(1 + \frac{\xi(x-\mu)}{\sigma} \right)^{-\frac{1}{\xi}} \right\} \left(1 + \frac{\xi(x-\mu)}{\sigma} \right)^{-\left(1+\frac{1}{\xi}\right)}. \quad (8)$$

Let $GEV(\mu, \sigma, \xi)$ be the GEV distribution with parameters μ , σ and ξ (Coles, 2001). A probabilistic model suitable for the cumulative range of a given day d is $GEV(\mu(t), \sigma(t), \xi)$. A regression model $GEV(\mu(t), \sigma(t), \xi)$ is then fitted to the set of observations $N_s(d)$. Extreme value model shape parameters are difficult to estimate with precision, so it is usually unrealistic to try modeling ξ as a smooth function of time. Because the range is limited from above, the maximum of any sample is asymptotically distributed according to an Extreme value distribution of the Weibull type (Type III) (Coles, 2001), i.e., a GEV distribution with $\xi < 0$. In this case it is also possible to make inferences at the upper end-point of the distribution, which is effectively the “infinite-observation return period”. The maximum likelihood estimate is $\hat{\mu} - \hat{\sigma}/\hat{\xi}$ where “^” denotes the point estimation. In practice, the model can be estimated using for instance, the function “gev.fit” in the R-package “ismev” (Heffernan, 2014). This function provides point estimations (and their respective standard errors) of the parameters of the best fit for the regression model $GEV(\mu(t), \sigma(t), \xi)$.

2.3.2 Control chart based on non-parametric methods

The first method, labelled *NPI*, is inspired by the Shewhart control charts. The control limits are set by adding to and subtracting from its average a multiple of the standard deviation of the controlled characteristic. In our case, the point-wise mean and the point-

wise standard deviation $m(d)$ and $sd(d)$ of the cumulative ranges in the set $N_s(d)$ of historical data, allow us to define the upper control limit curve in Eq.(9) as:

$$UCL1(d)=m(d)+\alpha_1*sd(d), \quad (9)$$

where α_1 determines the probability of a type I error. The usual value of α_1 in Shewhart control charts is three.

The second method, labelled *NP2*, is based on the covering set of all observed curves $N_s(d)$. The point-wise minimum and the point-wise maximum curves, $min(d)$ and $max(d)$ are obtained from the set $N_s(d)$, and the UCL curve is defined in Eq. (10) as:

$$UCL2(d)=max(d)+\alpha_2*sd(d), \quad (10)$$

where α_2 determines the probability of a type I error.

The third method, labelled *NP3*, is inspired by the boxplot to determine outliers. The first and third point-wise quartiles of the observations in $N_s(d)$, $Q1(d)$ and $Q3(d)$, are calculated. The UCL curve is defined in Eq. (11) as:

$$UCL3(d)=Q3(d)+\alpha_3*(Q3(d)-Q1(d)). \quad (11)$$

where α_3 determines the probability of a type I error. The usual value of α_3 in boxplot is 1.5.

The problem with these non-parametric approaches lies in justifying the choice of the parameters to determine the probability of a type I error. We could experiment with values for these parameters by means of simulation but it is not possible to derive generic results. That is why we will use these methods as descriptive methods in addition to the parametric method based on the GEV distribution.

The production of the PV panel sets can be monitored online. At time t of day d_0 , $X_k(t, d_0)$ is observed and the range $R(t, d_0)$ and the cumulative range $C(t, d_0)$ are calculated. The values of $C(d_0)$ are plotted in all control charts obtained from $N_s(d_0)$ (Eqs. (9)-(11)). $C(t, d_0)$ is then checked against $UCL(t, d_0)$ for every time index t . If $C(t, d_0)$ is larger than $UCL(t, d_0)$ for some time index t , there is an out of control signal that should be investigated. The control charts are refreshed each day and the monitoring restarts each day. As a result, the monitoring of the current day has no influence on what happened in the previous day.

3. Results and discussion

We applied our two-stage method as developed in the previous section to PV panel data of Flukso. The data we used as input in our first stage relate to a PV panel installation on one roof in Belgium. It is worth to notice that our method is developed to detect all kinds of outliers that affect the generated number of kWh of PV arrays. That is one of the reasons why we have not induced specific fault types in this experimental set-up. Another reason relates to the PV panel working of this real-life application which could not be interrupted (so we were not allowed to use it for a test setting). Moreover, as the real challenge is to detect small changes in energy output, we simulated a 5% reduction in power production by reducing the recorded production of one randomly chosen inverter by 5% in Section 3.2.

The installation consists of six sets of rack-mounted solar panels, with each set coupled to an inverter with two or three arrays. The sets of PV panels have the same orientation and inclination slope. More specifically, the roof contains 180 identical panels for which 6 inverters are used. The first three panel sets (arrays), each connected to an inverter of type SMA-SMC 9000TL all contain three parallel strings of 12 panels (i.e., 36 panels per array). The last three panels sets (arrays), each connected to an inverter of type SMA-SMC 6000TL all contain two parallel strings of 12 panels (i.e., 24 panels per array). This can be seen in Figure 2 where the panel arrays with three parallel strings of 12 panels are indicated by the symbols D (strings D1, D2 and D3), E (strings E1, E2 and E3) and F (strings F1, F2 and F3); the panel arrays with two parallel strings of 12 panels are indicated by the symbols A (strings A1 and A2), B (strings B1 and B2) and C (strings C1 and C2). So note that the energy production of the last 3 panel sets is 6/9 of the energy production of the first 3 sets. The DC electrical characteristics of the polycrystalline silicon panels can be found in Table 2. The data relate to the production of three different years for a total of 992 days: from 12/03/2011 until 31/12/2011 (295 days), from 01/01/2012 until 23/04/2012 and from 02/05/2012 until 31/12/2012 (357 days) and from 01/01/2013 until 06/12/2013 (340 days). The data represent the generated number of kWh (corresponding to PV panel DC output, also called P_{DC} or PDC value in what follows) for each 5 minute interval for each set of PV panels for the listed days. We now discuss respectively the first stage and second stage (see Figure 1) methods applied to our real-life case in Sections 3.1 and 3.2 followed by a discussion of the developed methodology in Section 3.3.

Figure 2. PV configuration on a roof in Belgium

DC electrical characteristic	Value
STC Power Rating	280W
PTC Power Rating	250.8W (California Solar Initiative (CSI) list of Eligible Modules)
STC Power per unit of area	13.4W/ft ² (144.3W/m ²)
Peak Efficiency	14.4%
Number of Cells	72
Nominal Voltage	not applicable
Imp	7.63A

V _{mp}	36.72V
I _{sc}	7.98A
V _{oc}	43.8V
Temp. Coefficient of Power	-0.46%/K
Temp. Coefficient of Voltage	-0.151V/K
Series Fuse Rating	12A
Maximum System Voltage	600V

Table 2. The DC electrical characteristics of each PV panels.

3.1 First stage results

We first remove those days containing a very small number of DC output measurements. Only March 14th, 2011 is removed as it contains 28 records. A representation of the number of measurements (i.e., number of 5 minute intervals with a P_{DC} measurement) of energy production records per day after removing March 14th, 2011 is represented in Figure 3. Note that the influence of the season can be clearly observed.

Figure 3. Number of measurements with recorded energy production per day for 992 days. Day number 1 is 12/03/2011.

In the following step, we rescale the energy production data of PV panel sets to bring them at the same level. This rescaling is needed because certain sets of PV panels have a different capacity. The energy production of the first three sets is divided by 9 whereas the production of the other sets is divided by 6 because the energy production capacity of the last three sets (i.e., panel sets 4, 5 and 6) is 6/9 of the energy production capacity of the first three sets (i.e., panel sets 1, 2 and 3). The results without and with rescaling for a specific day (May 15th, 2013) can be observed in Figure 4.

Figure 4. P_{DC} value as a function of time for two panel sets with different capacities without (left) and with rescaling (right) for May 15th, 2013.

The daily energy production is then stored in vectors $\{X_k(t, d); t = 1, \dots, 288; k = 1, \dots, 6\}_{d=1}^{992}$ with a length of 288 (1440 minutes/day divided by 5 minutes). For 5 minute intervals with no production, we assume a value of zero. Next, for each day, we obtain the range $R(d)$ and the cumulative range $C(d)$ as defined in Eqs. (2) and (4). A representation of the six energy production vectors for July 24th, 2013 and their associated range and cumulative range can be found in Figure 5 together with a representation of the cumulative ranges for all days in July 2013.

Figure 5. Representation of the P_{DC} value (left) and the range $R(d)$ (center) as a function of time for each panel set for July 24th 2013; a representation of the cumulative ranges $C(d)$ (right) for all days in July 2013 (the cumulative range corresponding to July 24th 2013 is highlighted).

Next, we construct the boxplots of the $L(d)$ values per year (see Eq. (6)) to detect global outliers. The largest norms correspond to the malfunctioning of at least one set of PV panels on particular days. This malfunctioning on specific days is easy to detect as the cumulative ranges for those days differ significantly from the normal cumulative ranges. The boxplots of the $L(d)$ values are presented in Figure 6.

Figure 6. Boxplots of the $L(d)$ values

From visual inspection of the boxplots of Figure 6, we find two types of outliers. The first type corresponds to outliers whose $L(d)$ value is larger than 8000 (dotted line). Those outliers represent days on which at least one set of PV panels functions abnormally. These days are: from May 30th to June 23rd (in 2011), from March 21st to April 2nd and from October 27th to November 11th (in 2012) and one day in January 2013. The second type of outliers (with $L(d)$ values smaller than 8000) has a very different explanation. As we will see, the explanation follows from applying FPCA to the range data $R(d)$ which makes it possible to detect a shadow pattern during the autumn-winter period which is not present in the spring-summer period. All days corresponding to the first type of outliers are removed before performing the FPCA.

The first step in applying FPCA is to smooth the discrete range vectors (Eq. (2)). Smoothed functional data are then created by fitting a B-spline basis to the set of range vectors $\{R(d)\}_{d=1}^{992}$. More specifically, we suggest using an order four B-spline basis function expansion with knots at sampling points. The application of the Generalized-Cross-Validation criterion suggests that a basis with 80 until 90 functions will be appropriate as can be observed from Figure 7. Therefore, we will opt for a value of 86.

Figure 7. Generalized-Cross-Validation scores as a function of the number of basis functions in applying FDA

We can now apply FPCA to the smoothed ranges. Figure 8 (top left) shows the first three principal components. However the interpretation of these components is not straightforward. Therefore, we will examine plots of the overall mean function together with the functions obtained by adding and subtracting a suitable multiple of each principal component function and the boxplot of the scores. This process is illustrated for the first principal component in Figure 8 (top right). The first, second and third principal components account for 43%, 15.2% and 10.1% of the variance respectively. We will focus on the first PC which accounts for the largest variability.

Figure 8. Representation of the first three PCs (top left) and of the overall mean function as well as two functions obtained by adding (+) and subtracting (-) a multiple of the first PC function (top right); the representation of the rescaled production of the same two panel sets on February 10th 2013 (bottom left) and June 12th 2013 (bottom center) together with their raw (non-smoothed) ranges (bottom right).

Figure 8 (top right) suggests that days with high scores for the first principal component have a range that is larger than the mean between approximately 1pm and 4pm. Besides, the first PC scores for each day are displayed in Figure 9 (left). We clearly observe the presence of a seasonal component. In Figure 9 (right) we present a boxplot of the first principal component scores. The days having the largest scores correspond to sunny days between the second part of October and March, e.g., February 10th 2013 (Figure 8 (bottom right)). This behavior corresponds to panels shadowing each other on sunny days due to the low position of the sun during this period. As such they receive less than normal solar

radiation. Note that as a consequence of this shadowing effect, the variability in the neighborhoods defined for autumn and winter days could be larger than the variability in the neighborhoods defined for spring and summer days (it depends on the proportion of sunny days in the autumn or winter day neighborhood). A representative spring day is e.g., June 12th 2013 whose range is represented in Figure 8 (bottom right). The energy production of the different panel sets during such type of days is not only higher than in autumn and winter but also more similar because there is no “shadowing effect” (indeed, the difference between panel sets 5 and 6 is indistinguishable for June 12th 2013 in Figure 8 (bottom center)).

Figure 9. First principal component scores by day (left) and boxplot of the first principal component scores (right).

3.2 Second stage results

In defining the daily neighborhoods $N_s(d)$, we fix s at seven, and (after removing “out of control” data) we have enough data to take into account two years. This choice implies that each neighborhood comprises 30 observations ($n(d)=(1+2*s)*2=30$). We now present different examples based on the “in control state” data resulting from Section 3.1 to illustrate the performance of the control charts.

In Figure 10, we first represent the neighborhood for October 1st (an autumn day with high variability in the neighborhood because, although not as intense as it happens later in the autumn and winter period, the “shadowing effect” starts to appear) as well as for July 24th (a sunny summer day with low variability in the neighborhood). Indeed, in Figure 10 (left) we observe that for some days, at approximately 5pm (=minute 1000), there is a “fast increment” in the cumulative range which corresponds to the “shadowing effect”.

Figure 10. Representation of the cumulative ranges for the neighborhood days of October 1st (left) and July 24th (right). Their respective point-wise maximum, minimum, mean and quartiles are also represented.

In Figure 11 we present the UCLs of the three non-parametric methods $NP1$, $NP2$ and $NP3$ (see Eqs. (9)-(11)) as a black line for October 1st and July 24th.

Figure 11. Cumulative ranges of the neighborhood and UCL_{NP1} (top left), UCL_{NP2} (top center) and UCL_{NP3} (top right) for July 24th and UCL_{NP1} (bottom left), UCL_{NP2} (bottom center) and UCL_{NP3} (bottom right) for October 1st.

The construction of the upper limit using the GEV regression method (UL_{GEV} or UL_{GEV}) is illustrated for July 24th (see also Section 2.3.1). The parametric frontier based on the GEV distribution is computed by fitting a model to its neighborhood. We select the time indices corresponding to 10am to 5pm, i.e., from time index t equal to 121 till 204. That is because we want to remove the part of the day in which the energy production is null or less significant but with an assignable cause of variability: the differences in the number of solar hours among the days included in the neighborhood. This allows us to fit a GEV. For computational reasons, the time indices are standardized and 0 corresponds to 1:30pm. For simplicity we will also use the term “ranges” to represent these reduced and time standardized versions of the original ranges. Instead of the usual “type-I graphic”, in Figure 12 we represent the observations as points to emphasize that we are fitting a regression model. We now consider each cumulative range as a realization of a non-stationary process. The $GEV(\mu(t), \sigma(t), \xi)$ distribution fitted to the 24th of July neighborhood data has the following expression for its parameters

$$\mu(t) = \alpha_\mu + \beta_\mu * t = 919.9 + 391.2 * t$$

$$\sigma(t) = \alpha_\sigma + \beta_\sigma * t = 247.9 + 106.1 * t$$

$$\xi = -0.24,$$

where the standard errors of the parameters $\alpha_\mu, \beta_\mu, \alpha_\sigma, \beta_\sigma, \xi$ are respectively: 5.8, 4.5, 3.9, 2.7 and 0.011. For July 24th, we compute a time dependent upper limit $UL_{GEV}(t)$ as:

$$UL_{GEV}(t) = 919.8 + 391.2 * t + (247.9 + 106.1 * t) / 0.24.$$

In Figure 12 the upper limit ($UL_{GEV}(t)$) is represented as a continuous magenta line.

Figure 12. Cumulative ranges of the neighborhood and UL_{GEV} (solid line) for July 24th.

In order to check whether the three non-parametric methods allow for the detection of small failures, we build a test case where one of the sets of PV panels has an energy production equal to 0.9 times the initial energy production. We clearly observe in Figure 13 that the

cumulative range goes out of the control zone for all methods on October 1st as well as July 24th.

Figure 13. Cumulative ranges of the neighborhood (grey curves), UCL_{NP1} , UCL_{NP2} and UCL_{NP3} (colored solid lines) and a test curve corresponding to one set of panels having a production value scaled with a factor of 0.9 for October 1st (left) and for July 24th (right) (black dotted curve).

All methods (also the parametric method) perform well. Even in cases where the reduction in the efficiency of one panel set is only 5% it is possible to receive an out of control signal, as can be seen in Figure 14 where the production of July 24th is reduced to 95% of its initial value.

Figure 14. Cumulative ranges of the neighborhood (grey curves) and UCL_{NP1} , UCL_{NP2} and UCL_{NP3} (colored solid lines) (left) and UL_{GEV} (magenta line) and a test curve corresponding to one set of panels having a production value scaled with a factor of 0.95 (black dotted curve) and the neighborhood of July 24th (right).

3.3 Discussion

Our new automatic method is developed to detect all kinds of outliers that affect the generated number of kWh in PV systems consisting of identical sets (sister arrays). It has been tested on real data obtained from monitoring PV panels in real conditions of work. We have shown that this new method is able to detect the malfunctioning of one or more sets of PV panels for very small changes in production output (i.e., a 5% reduction in power production of one of the sets). However, such small changes might render it difficult to distinguish true faults (which require intervention) from normal operating panels. Indeed some conditions might generate error detection while no true faults are occurring. On top, the type I error also causes detecting unjustified errors. In order not to intervene too rapidly, we propose to implement the following procedure once an outlier is detected. We first check if the outlier detection lasts for at least a specific period (e.g., 1 day). At the same time, we propose to simultaneously keep track of the scaled (to bring the energy production at the same level for all sets) generated number of kWh (i.e., the PDC value) for each minute interval of each set of PV panels. If the outlier detection still lasts after this period and the scaled generated number of kWh from one or more panel sets indeed shows that there is a loss in power, we will intervene in the solar panel system. However, as deriving a robust value for the length of the error detection period before considering this detection as a true error, requires additional extensive analysis and thorough testing, it is considered outside the scope of our paper and taken as point for future research. Note that because faults in the system induce high ranges, they can be easily detected by our methodology and were not part of our experimental set-up in Section 3.2.

The method is based only on a comparative analysis of the monitored PV panel yield production. On the one hand, this avoids the development of simulation models as well as the gathering of climate data to feed them, as is necessary in most of the methods presented in the literature. Indeed, in those methods the monitoring of real data is complemented with the monitoring of some climate data that are obtained either by sensors at the PV system location or from satellite observations. The fault detection algorithm is then based on the comparison of simulated and measured productions by analyzing the losses in the PV system (see e. g., Drews et al., 2007; Chouder and Silvestre, 2010; Silvestre et al., 2013; Tadj et al., 2014). The data necessary to apply our method are thus simpler to obtain making it an attractive alternative method of monitoring energy production. To the best of our knowledge, Zhao et al. (2013) are the only ones that also perform a comparative analysis by relying on differences between measurements from adjacent strings for fault detection in PV strings. However, they do not derive “in control” data whereas we believe that this “in control” reference dataset is an important part of the methodology to get reliable error detection. So our method detects outliers in the range of current production which, once the control chart is calibrated, does not need any extra information. On the other hand, this comparative analysis asks for sets that are identical with respect to all characteristics (i.e., location, orientation, inclination but also age, module material, BOS components, and local thermal environment, etc.) in order for our method to be reliable. Differences in these characteristics will indeed reduce correlation between the sister arrays. That is also the reason why our model is in the first place suited for PV “farms” (as in the real-life example given in Section 3). However, we believe that our method could also be extended to non-identical arrays by “aligning” the production data of the different arrays using techniques

as regression analysis. More specifically, to apply our method to production data of arrays with different characteristics, an extra pre-processing phase should be added to our approach in order to statistically correct the production data for (measurable) differences in the array characteristics. The pre-processing phase should be performed two times: (1) to construct the “in control” reference data set as well as (2) to compare the real data with the “in control” dataset. Assuming that the characteristics of the arrays can be quantified with a number of variables (e.g., orientation, inclination, age of the array, module material, etc.), historic production data can be corrected for array differences by conducting a regression analysis of production data on array characteristics. If the regression model is valid (i.e., the assumptions of the model are not violated), then the residuals of this regression analysis can be considered as production data which are corrected for possible array differences, and which are therefore comparable across arrays. However, as this preprocessing requires an additional in depth pre-processing phase to construct reliable “in control” data and transforming the real data to make them comparable with the “in control” data, it is outside the scope of our paper. Note also that our method is not able to pinpoint the specific panel that is malfunctioning as it operates on the relative difference in energy production of sets of panels.

We would like to comment that, some of the methods which monitor the differences between real and simulated production (see e. g., Chouder and Silvestre, 2010; Silvestre et al., 2013), set control limits inspired by the Shewhart control charts in the sense that they estimate the expected value and set the threshold at two times the standard deviation. One of our non-parametric methods follows the same strategy although we recommend, as is common in quality control charts, counting three standard deviations from the expected value.

A key issue in our method is the definition of the neighborhood. In the implementation of the method the neighborhood has been built in a simple way by considering the similarity only by means of proximity in the calendar. Although this is a reasonable approach, the definition of the neighborhood could be improved by considering not only the calendar proximity but also the type of day. For example, each day could have several neighborhoods depending on the type of day. We could consider four types of days: “Clear sky days”, “Intermittent clear sky days”, “Intermittent cloudy sky days” and “Cloudy sky days”. If a given day is expected to be, let us say, a clear sky day, then the curves in its neighborhood which are close in time should correspond to clear sky days. A method to classify the days in types according to their solar radiation curves is provided in Gaston et al. (2011). They use functional data analysis and mathematical morphology to classify days according to radiation levels. This method could be used prior to the construction of the daily neighborhoods. This modification of the definition of neighborhood would reduce the variability of the corresponding cumulative ranges, although more background data would be required.

4. Conclusion

A new method has been presented for automatically detecting outliers or faults in solar energy production among identical sets (sister arrays) of PV panels (i.e., sets sharing the same characteristics such as location, orientation, inclination slope, age, module material, BOS components, local thermal environment, etc.). Note that as the differences in the sister

arrays will affect the effectiveness of our method, the approach we have developed in this paper should operate on production data from identical sets. To detect differences in the energy production of these sets of panels (caused by outliers affecting the generated number of kWh) the method operates on the distribution of the cumulative range.

The presented method involves an unsupervised two-stage approach. The first stage consists of creating “in control state” data by removing days with abnormal behavior. More specifically, after removing days with very few observations, a boxplot of the L2-norm values of the cumulative ranges is used to remove days with global malfunctioning. Next, we apply functional principal components analysis to detect days with local temporal deviations (e.g., shadowing) in the production data. In the second stage, quality control charts for monitoring the energy production of a new day are constructed using the sample of “in control” data. For each day a neighborhood of days is composed in which the range and cumulative range are expected to be statistically similar. Next, in order to classify new observations as being generated by the “in control” or “out-of-control” energy production process, different types of upper (control) limits for the cumulative range are constructed. In particular, we propose a parametric model that uses extreme value distributions to estimate the upper limit of the cumulative range distribution in the “in control” state and three non-parametric methods.

As an illustration, the method is applied to analyzing real energy production data of six sets of PV panels over a period of three years. Tests indicate that the proposed method is able to successfully detect very small changes in energy output, i.e., a reduction in efficiency in one of the solar panel sets by up to 5%. Control charts based on the developed parametric and non-parametric methods performed well. However, small changes in energy output might render it difficult to distinguish true faults from normal operating panels. So it is appropriate to check if the outlier error detection lasts for a specific period in order not to intervene too rapidly. A thorough analysis to determine a robust value for the length of this period would therefore be an interesting topic for future research.

As the proposed method only relies on historical data, it could easily be extended in order to detect outliers in the electricity use of household appliances, water or gas usage by considering, for example, each year of available data as a data set. Moreover, we believe that our method could also be extended to non-identical sets (arrays) by “aligning” the production data of the different arrays using techniques as regression analysis. More specifically, to apply our method to production data of arrays with different characteristics, an extra pre-processing phase should be added to our approach in order to statistically correct the production data for (measurable) differences in the array characteristics.

5. Acknowledgments

This paper has been supported under Grants MTM 2012-36025 and MTM2016-77015-R.

References

- Akram, M.N., Lotfifard, S., 2015. Modeling and Health Monitoring of DC Side of Photovoltaic Array. *IEEE Transactions on Sustainable Energy*, 6 (4), 1245-1253.
- Alam, M.K., Khan, F., Johnson, J., Flicker, J.A., 2015. Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques. *IEEE Journal of Photovoltaics* 5 (3), 982-997.

- Aguinis, H., Gottfredson, R.K., Joo, H., 2013. Best-practice recommendations for defining, identifying and handling outliers. *Organizational Research Methods* 16(2), 270-301.
- Capozzoli, A., Lauro, F., Khan, I., 2015. Fault detection analysis using data mining techniques for a cluster of smart office buildings. *Expert Systems with Applications* 42, 4324-4338.
- Ceroli, A., Laurini, F., Corbellini, A., 2005. Functional cluster analysis of financial timeseries, in: Bock, H.H., Gaul, W., Vichi, M. (Eds.), *New developments in Classification and Data Analysis, Proceedings of the meeting of the CLADAG of the Italian Statistics Society, University of Bologna, September 22-24, 2003. Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 333-341). Heidelberg, Berlin: Springer.
- Chine, W., Mellit, A., Pavan, A.M., Kalogirou, S.A., 2014. Fault detection method for grid-connected photovoltaic plants. *Renewable Energy* 66, 99-110.
- Chouder, A., Silvestre, S., 2010. Automatic supervision and fault detection of PV systems based on power losses analysis. *Energy Conversion and Management* 51, 1929-1937.
- Chouder, A., Silvestre, S., Taghezouit, B., Karatepe, E., 2013. Monitoring, modeling and simulation of PV systems using LabVIEW. *Solar Energy* 91, 337-349.
- Coles, S., 2001. *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.
- De Boeck, L., Van Asch, S., de Bruecker, P., Audenaert, A., 2016. Comparison of support policies for residential photovoltaic systems in the major EU markets through investment profitability. *Renewable Energy* 87(1), 42-53.
- Dodier, R.H., Kreider, J.F., 1999. Detecting whole building energy problems. *ASHRAE Transactions* 105, 579-589.
- Drews, A., de Keizer, A.C., Beyer, H.G., Lorenz, E., Betcke, J., van Sark W.G.J.H.M., Heydenreich, W., Wiemken, E., Stettler, S., Toggweiler, P., Bofinger, S., Schneider, M., Heilscher, G., Heinemann, D., 2007. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. *Solar Energy* 81, 548-64.
- European Commission, 2007. The 2020 climate and energy package. Retrieved January 5, 2014, from http://ec.europa.eu/clima/policies/package/index_en.htm.
- Firth, S.K., Lomas, K.J., Rees, S.J., 2010. A simple model of PV system performance and its use in fault detection. *Solar Energy* 84, 624-635.
- Gastón, M., León, T., Mallor, F., Ramírez, L., 2011. A morphological clustering method for daily solar radiation curves. *Solar Energy* 85, 1824-1836.
- Gokmen, N., Karatepe, E., Celik, B., Silvestre, S., 2012. Simple diagnostic approach for determining of faulted PV modules in string based PV arrays. *Solar Energy* 86, 3364-3377.

Heffernan, J. E., Stephenson, A.G., 2014. ismev: An Introduction to Statistical Modeling of Extreme Values. <http://CRAN.R-project.org/package=ismev>

Hu, Y., Gao, B., Song, X., Tiang, G.Y., Li, K., He, X., 2013. Photovoltaic fault detection using a parameter based model. *Solar Energy* 96, 96–102.

Platon, R., Martel, J., Woodruff, N., Chau, T.Y., 2015. Online Fault Detection in PV Systems. *IEEE Transactions on Sustainable Energy* 6(4), 1200–1207.

Quesada, B., Sánchez, C., Canada, J., Royo, R., Payá, J., 2011. Experimental results and simulation with TRNSYS of a 7.2 kWp grid-connected photovoltaic system. *Applied Energy* 88, 1772–1783.

Ramsay, J.O., Silverman, B.W., 2005. *Functional Data Analysis*, second ed. Springer, New York.

Ramsay, J.O., Wickham, H., Graves, S., Hooker, G., 2014. fda: Functional Data Analysis. R package version 2.4.4. <http://CRAN.R-project.org/package=fda>

Silvestre, S., Chouder, A., Karatepe, E., 2013. Automatic fault detection in grid connected PV systems. *Solar Energy* 94, 119–127.

Silvestre, S., da Silva, M.A., Chouder, A., Guasch, D., Karatepe, E., 2014. New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators. *Energy Conversion and Management* 86, 241–249.

Silvestre, S., Kichou, S., Chouder, A., Nofuentes, G., Karatepe, E., 2015. Analysis of current and voltage indicators in grid connected PV (photovoltaic) systems working in faulty and partial shading conditions. *Energy* 86, 42–50.

Stettler, S., Toggweiler, P., Remund, J., 2006. SPYCE: satellite photovoltaic yield control and evaluation, in: *Proceedings of the 21st European photovoltaic solar energy conference*, pp. 2613–2616.

Tadj, M., Benmouiza, K., Cheknane, A., Silvestre, S., 2014. Improving the performance of PV systems by faults detection using GISTEL approach. *Energy Conversion and Management* 80, 298–304.

Vatanen, T., Kuusela, M., Raiko, T., Aaltonen, T., Nagai, Y., 2012. Semi-Supervised Detection of Collective Anomalies with an application in high energy particle physics. *WCCI 2012 IEEE World Congress on Computational Intelligence*.

Vergura, S., Acciani, G., Amoroso, V., Patrono, G.E., & Vacca, F., 2009. Descriptive and Inferential Statistics for Supervising and Monitoring the Operation of PV Plants. *IEEE Transactions on Industrial Electronics* 56(11), 4456–4464.

Vergura, S., Acciani, G., Falcone, O., 2012. A finite-element approach to analyze the thermal effect of defects on silicon-based PV cells. *Transactions on Industrial Electronics* 59(10), 3860–3867.

Xiaoli, L., Bowers, C.P., Schnier, T., 2010. Classification of energy consumption in buildings with outlier detection. *IEEE transactions on industrial electronics* 57(11), 3639-3644.

Zhao, Y., Lehman, B., Ball, R., Mosessian, J., de Palma, J., 2013. Outlier detection rules for fault detection in solar photovoltaic arrays, in: *Proceeding of 28th Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition*, pp. 2913–2920.