NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Anomaly detection in photovoltaic systems

Pedro Miguel Mayer Branco

Internship report presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ANOMALY DETECTION IN PHOTOVOLTAIC SYSTEMS

by

Pedro Miguel Mayer Branco

Internship report presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Information Analysis and Management

Advisor: Dr. Ana Cristina Costa

Co-advisor*:* Eng. Francisco Gonçalves

March 2019

ABSTRACT

Photovoltaic (PV) solar energy is the fastest-growing renewable source of energy, and poised to become the world's largest source of electricity by 2050. To maximize efficiency and remain a viable alternative energy source, PV systems should ideally operate seamlessly without anomalies. In reality, however, several kinds of anomalies may occur that prevent PV systems from operating at their full capacity. Here, we address this problem by developing five algorithms for the detection of several PV-system anomalies, and establishing metrics to determine the severity of daytime shading and suboptimal orientation. Specifically, our algorithms are used to detect brief and sustained daytime zero-production, daytime and sunrise/sunset shading, low maximum production and suboptimal orientation. We apply these detection algorithms to several time-series of electricity production, which were obtained for two periods with contrasting weather conditions. When weather conditions were favorable, our algorithms successfully detected the majority of time-series labeled with either sustained or brief daytime zero-production, and either daytime or sunrise/sunset shading. Furthermore, these algorithms also produced a relatively low percentage of false positives, which indicates that most anomaly detections are correct. When weather conditions were adverse, the detection rate of our algorithms was similarly high, if not higher, than when weather conditions were favorable. However, the percentage of false positive anomaly detections is also substantially higher under adverse weather conditions, which indicates that the algorithms are generally more robust under favorable weather conditions. Our results suggest that, on the one hand, daytime shading is a relatively rare anomaly, although it may have a severe impact on PV-system efficiency that warrants its detection. On the other hand, suboptimal orientation appears to be relatively common, and our orientation index can therefore be useful to determine the severity of this prevalent type of anomaly.

KEYWORDS

Anomaly detection; photovoltaic systems; renewable energy; time-series

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The concentration of greenhouse gases (GHGs) in Earth's atmosphere has been steadily increasing since the Industrial Revolution (Solomon et al., 2007), and thereby contributing to global warming over the last two centuries (Houghton et al., 2002; Lashof & Ahuja, 1990; Meinshausen et al., 2009). This trend is largely due to human activity, where the energy sector has become a major driver of GHG emissions (IEA, 2014a). Indeed, if current practices in energy production are to be maintained, global temperature is expected to increase as much as 6 °C above pre-industrial levels by 2050 (IEA, 2014b). To meet the emissions target yielding global warming of 1.5 °C by 2050 (UNFCCC, 2018), there is therefore an urgent need to shift from carbon-intensive to greener and renewable energy sources.

Photovoltaic (PV) solar energy is perhaps the most promising and fastest-growing renewable source of energy, having generated 1.7% of global power in 2017 (BP, 2018) and poised to become the world's largest source of electricity by 2050 (IEA, 2014a). Hence, PV systems have great potential to reduce our current dependence on carbon-intensive sources of energy, and progress has been made in enhancing their efficiency. In fact, solar cell efficiency has increased from about 5% to over 40% over the last 60 years (Dimroth et al., 2014), yet current efficiency levels are rather low compared to those of alternative sources of energy.

To maximize efficiency and remain a viable alternative energy source, PV systems should ideally operate seamlessly without anomalies. In reality, however, several kinds of anomalies may occur that prevent PV systems from operating at their full capacity. Thus, it is important to monitor the activity of PV systems, so that these anomalies can be detected and repaired to ensure maximum efficiency (Chouder & Silvestre, 2010; Drews et al., 2007; Firth et al., 2010; Platon et al., 2015). Here, we address this problem by developing algorithms for anomaly detection in PV systems, and establishing metrics to estimate anomaly severity.

1.2. STUDY OBJECTIVES

The main objective of this internship is to develop tools that automatically detect anomalies in the electricity production of PV systems. In particular, we aim to develop detection algorithms indicating whether and why a given PV system is operating anomalously, and establish simple metrics to estimate anomaly severity. This approach will allow us to distinguish between several types of production anomaly (e.g., inverter shutdown, shading), and thereby tackle anomalies in the most efficient manner. We refer to section 2 for a brief literature review on anomaly detection techniques, and section 3 for a detailed description of the methodology used in our work.

1.3. THE COMPANY AND THE DATASET

This internship was carried out at CSide (www.cside.pt). CSide is a software company specialized in the management of energy production and consumption in residential and commercial buildings.

More specifically, this company develops digital applications for service providers, energy utilities and telecommunication companies, which inform end-users about their energy production and consumption patterns. In this way, CSide seeks to raise awareness of energy production and consumption, and thus promote energy-savvy habits among end-users.

Our research goal of automatically detecting anomalies in PV systems requires access to data on electricity production, so that electricity production levels can be assessed over time. To this end, CSide holds a database including 1676 univariate time-series of electricity production. Each time-series corresponds to a different PV system, and consists of electricity production measurements recorded every 15 minutes over several months to years (Fig. 1.1). We use this dataset to develop and test anomaly detection algorithms, as well as anomaly severity metrics.

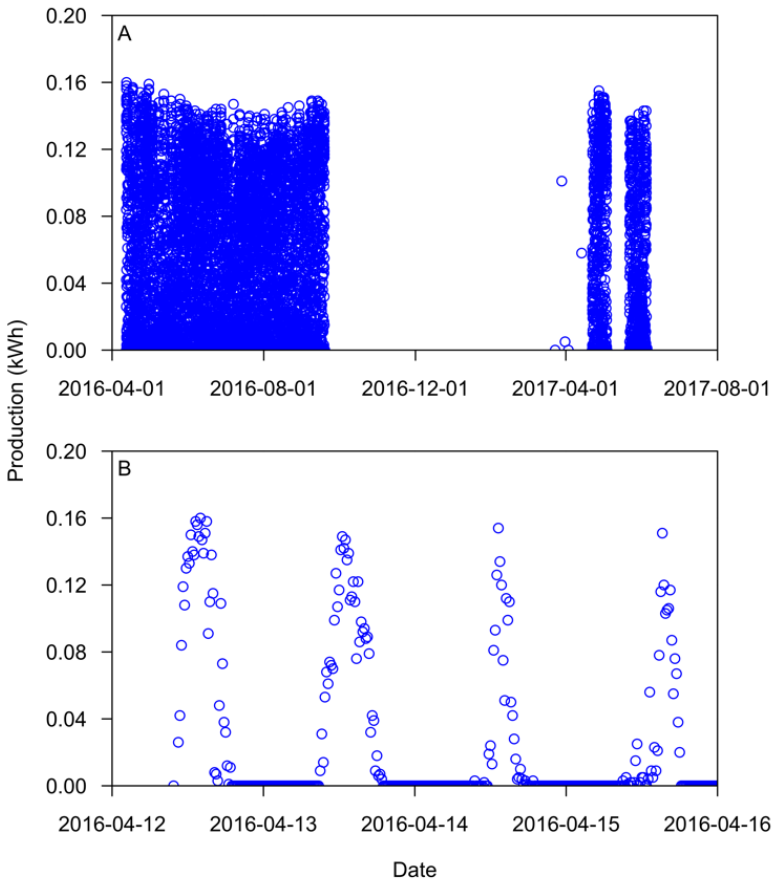

Figure 1.1 – Example of a time-series of electricity production analyzed in this study. *A*, The complete time-series. *B*, An excerpt from the complete time-series showing the first four days of electricity production.

2. LITERATURE REVIEW

2.1. ANOMALY DEFINITION AND CLASSIFICATION

Anomalies, also referred to as outliers, can be defined as observations that seem inconsistent with the dataset to which they belong (Barnett & Lewis, 1994). Anomalous observations therefore deviate from the norm, and occur due to a variety of reasons. For example, anomalous observations may result from measurement errors; fraudulent behavior (Bolton & Hand, 2002); the emergence of a novel pattern (Markou & Singh, 2003a,b); or faulty system performance (Firth et al., 2010). Depending on their characteristics, anomalies can be classified into three basic types (Chandola et al., 2009):

- *Point anomalies*, which are observations that deviate from the rest of the dataset.
- *Contextual anomalies*, which are observations that deviate from their context in the dataset. That is, a contextual anomaly is an observation that is considered anomalous in its context, but could otherwise be considered normal in a different context of the dataset.
- *Collective anomalies*, which are collections of related observations that deviate from the rest of the dataset. While each of the observations in a collective anomaly may be considered normal on its own, their specific combination is considered anomalous.

These three types of anomalies are likely to occur in time-series, such as those analyzed in this study. Hence, in our dataset, point anomalies correspond to levels of electricity production outside the normal range of measurements; contextual anomalies correspond to levels of electricity production that are inconsistent at their time of occurrence; and collective anomalies correspond to sequences of higher or lower electricity production than expected.

2.2. ANOMALY DETECTION

Anomaly detection is the process of identifying inconsistent observations in datasets, and includes several methods ranging from classical statistics to data mining and machine learning (Chandola et al., 2009; Hodge & Austin, 2004). The choice of method for anomaly detection primarily depends on the degree of prior knowledge of the data. In particular, Chandola et al. (2009) and Hodge & Austin (2004) distinguish three fundamental methods of anomaly detection:

- *Supervised anomaly detection*, which requires a training dataset where both normal and anomalous observations are labeled. Typically, these methods build a predictive model during training, and use this model to classify unseen observations as normal or anomalous.
- *Semi-supervised anomaly detection*, which requires a training dataset where only normal observations are labeled. These methods are more widely used than those of supervised anomaly detection, because they do not require anomalous observations to be labeled in the training dataset.

- *Unsupervised anomaly detection*, which does not require a training dataset with labeled observations, and thus is most widely used. Importantly, these methods implicitly assume that normal observations far outnumber anomalous observations, and violating this assumption may lead to an increase in false anomalous classifications.

Our dataset consists of univariate time-series where normal and anomalous observations are not labeled, and we therefore focus our literature review on methods of unsupervised anomaly detection. Specifically, we consider the strengths and weaknesses of commonly used statistical methods, distance-based methods, and neural networks.

2.2.1. Statistical methods

2.2.1.1. Grubbs' method

Early methods of unsupervised anomaly detection are based on classical statistics, and their application is largely restricted to univariate and quantitative data (Hodge & Austin, 2004). Perhaps the earliest anomaly detection rule is Grubbs' method (Grubbs, 1950), which computes the standardized value Z_i of each observation x_i in a sample:

$$Z_i = \frac{x_i - \bar{x}}{s}, \quad (1)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, and $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 / (n-1)}$ is the standard deviation. Thus, observation x_i is classified as an anomaly if $|Z_i|$ exceeds a critical value, which corresponds to an arbitrary number of standard deviations away from the sample mean. Typically, observation x_i is considered anomalous if it is more than three standard deviations away from the sample mean (i.e., if $|Z_i| > 3$), and normal otherwise.

2.2.1.2. Robust Grubbs' method

Grubbs' method relies on estimators of location and scale that are scarcely robust, and can therefore easily fail to detect clear anomalies (Rousseeuw & Hubert, 2011). To obtain a more robust method of anomaly detection, we can replace the sample mean and standard deviation in Equation (1) by alternative estimators of location and scale. Specifically, the sample mean can be replaced by the more robust sample median:

$$\text{median}_{i=1}^{n}(x_i) = \begin{cases} x_{[(n+1)/2]} & \text{if } n \text{ is odd} \\ \left[x_{(n/2)} + x_{(n/2+1)}\right]/2 & \text{if } n \text{ is even} \end{cases}, \quad (2)$$

whereas the standard deviation can be replaced by the more robust median of all absolute deviations from the sample median, *MAD*:

$$MAD = \alpha \, \text{median}_{i=1}^{n}\left(\left|x_i - \text{median}_{j=1}^{n}(x_j)\right|\right), \quad (3)$$

where the constant α is a correction factor that makes *MAD* an unbiased estimator. Similar to Grubbs' method, we first calculate the standardized value m_i of each observation x_i in a sample:

$$m_i = \frac{x_i - \text{median}_{j=1}^{n}(x_j)}{MAD}, \quad (4)$$

and subsequently determine whether $|m_i|$ exceeds a critical value, in which case observation x_i is considered an anomaly.

2.2.1.3. Tukey's fences

A popular method of unsupervised anomaly detection in univariate data is the so-called Tukey's fences (Tukey, 1977), which are defined by the following interval:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)], \quad (5)$$

where Q_1 and Q_3 are the lower and upper sample quartiles, respectively, $Q_3 - Q_1$ is the interquartile range, and k is a nonnegative constant. Thus, observations lying within Tukey's fences are considered normal, whereas observations lying outside Tukey's fences are considered anomalies. Tukey (1977) proposes that two values of k be used to detect anomalies: (1) k = 1.5, yielding 'inner' fences outside which anomalies occur; and (2) k = 3, yielding 'outer' fences outside which severe anomalies occur. Although originally developed for anomaly detection in univariate data, Tukey's fences have also been successfully applied to multivariate data (Laurikkala et al., 2000).

2.2.1.4. Q-Q plot

Quantile-quantile plots (Q-Q plots) are a graphical tool that compares two probability distributions (Wilk & Gnanadesikan, 1968). Specifically, the quantiles of a sample distribution are plotted against those of an appropriate theoretical distribution, and anomalies are detected as sample quantiles substantially different from theoretical quantiles. Although Q-Q plots typically compare the quantiles of a sample distribution with those of a standardized normal distribution $N(0,1)$, they can actually be used with any theoretical distribution. Furthermore, Q-Q plots work best with univariate and bivariate data, and anomaly detection becomes more difficult with multivariate data.

2.2.2. Distance-based methods

Unlike the statistical methods described above, distance-based methods of anomaly detection do not make prior assumptions about data distributions, and tend to perform well on multivariate data (Hodge & Austin, 2004). For continuous data, distance-based methods typically use the Euclidean distance between observations x and y, D_{xy}, to measure their dissimilarity:

$$D_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \quad (6)$$

More specifically, distance-based methods assume that normal observations occur close to their nearest neighbors and form dense neighborhoods, whereas anomalous observations occur far from their nearest neighbors and form sparse neighborhoods (Chandola et al., 2009).

Knorr & Ng (1998) proposed a distance-based definition of anomaly, where an observation is anomalous if it lies farther than distance *D* from at least a fraction *p* of all observations in the dataset. Although this approach works well for multivariate data, it has two important drawbacks (Ramaswamy et al., 2000). First, it only distinguishes normal observations from anomalous observations, and therefore does not rank anomalies according to their severity. Second, its output largely depends on the choice of parameters *D* and *p*, which are user-defined.

Alternatively, Ramaswamy et al. (2000) developed the k-Nearest Neighbor (k-NN) method of anomaly detection, which allows anomalies to be ranked according to their severity. Given a k^{th} nearest neighbor and *n* anomalies to be ranked, an observation is anomalous if no more than $n - 1$ other observations in the dataset have longer distance to their k^{th} nearest neighbor, D^k. Hence, the k-NN method is able to rank each observation according to its distance D^k, where the top *n* observations are considered anomalies.

Despite being able to rank anomalies, the k-NN method also has its drawbacks. In particular, its output largely depends on the choice of parameter *k*, which is user-defined (Madsen, 2018). On the one hand, choosing too low a value of *k* may prevent the detection of close anomalies in an outlying cluster. On the other hand, choosing too high a value of *k* may prevent the detection of severe anomalies.

2.2.3. Neural networks

Although statistical and distance-based methods are well suited to detect point anomalies in univariate and multivariate data, respectively, these methods are less suitable to handle contextual and collective anomalies typically found in time-series data (Chandola et al., 2009). Alternatively, there are neural networks that can learn the normal features of time-series in an unsupervised manner (Långkvist et al., 2014), and thereby detect contextual and collective anomalies. Below we introduce two such neural networks: the autoencoder and recurrent neural networks (RNNs).

2.2.3.1. Autoencoder

The autoencoder is a neural network that reconstructs its input (Fig. 2.1). More specifically, the autoencoder includes two main components: the encoder and the decoder. The encoder maps the input, x, to a code in the hidden layer, *h*, whereas the decoder maps *h* to the output space, $\hat{x}$ (Pereira, 2018). In other words, the autoencoder can take a time-series as input, $x = \{x(1), x(2), \dots, x(t)\}$, and learn its approximation, $\hat{x} = \{\hat{x}(1), \hat{x}(2), \dots, \hat{x}(t)\}$, such that $\hat{x}$ is similar to x (Ng, 2011).

Originally developed for dimensionality reduction (Hinton & Salakhutdinov, 2006), the autoencoder has also been used for anomaly detection in time-series data (Malhotra et al., 2016). During the training phase, the autoencoder learns the normal features of time-series, provided that normal observations far outnumber anomalous observations. Subsequently, the autoencoder is expected to have difficulties reconstructing anomalous time-series, because their features deviate from those of

normal time-series. As a result, anomalous time-series are detected due to their higher reconstruction error relative to normal time-series.

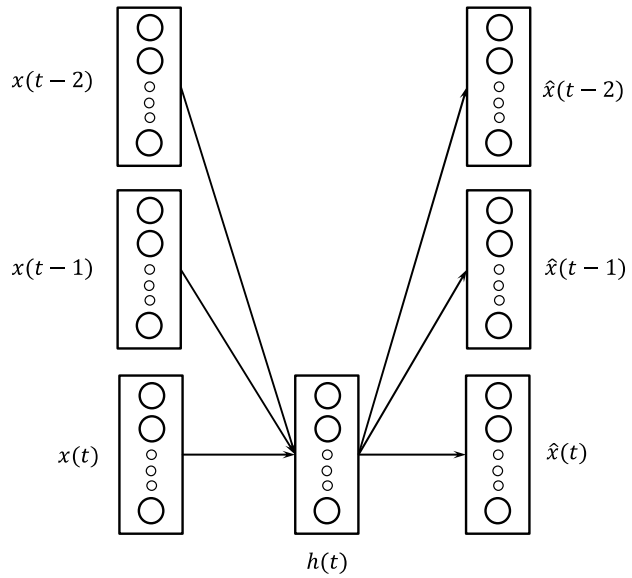

Figure 2.1 – Example of an autoencoder for time-series input, where x denotes the original time-series, *h* denotes the hidden layer, and $\hat{x}$ denotes the reconstruction of x. Figure adapted from Längkvist et al. (2014).

2.2.3.2. Recurrent neural networks

Although the autoencoder can be used for anomaly detection in time-series data, it does not capture the time dependencies that occur in the data. This problem is particularly pertinent when dealing with univariate data, where it is highly desirable to extract as much information from the data as possible (Pereira, 2018). RNNs solve this problem by using an architecture with interconnected hidden neurons (Fig. 2.2), and are therefore able to model short-term time dependencies (Hüsken & Stagge, 2003).

Long short-term memory (LSTM) networks are an extension of RNNs, which are able to find long-term time dependencies in the data (Hochreiter & Schmidhuber, 1997). LSTM networks have been successfully combined with autoencoders, yielding a network architecture that is suitable for anomaly detection in time-series (Malhotra et al., 2016). Indeed, such LSTM autoencoders are able to not only produce a representation of time-series data, but also take time dependencies into account. Hence, time-series reconstructions using LSTM autoencoders are more accurate than those using simple autoencoders, and anomaly detection is thereby improved.

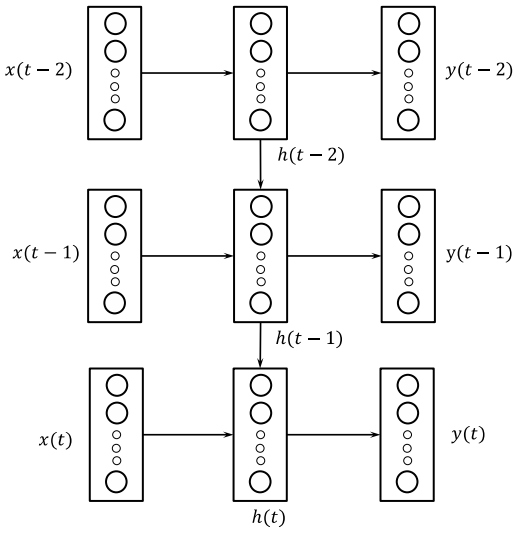

Figure 2.2 – Example of an RNN for time-series input, where x denotes the original time-series, *h* denotes the hidden layer, and y denotes the output representation. Figure adapted from Långkvist et al. (2014).

2.3. PV-SYSTEM ANOMALIES: AN OVERVIEW

Despite their potential to offer a clean and inexhaustible source of energy, PV systems often operate suboptimally due to several kinds of anomalies. We distinguish two major types of PV-system anomalies: (1) intrinsic PV-system faults, which originate from the PV system itself; and (2) extrinsic factors, which are external to the PV system and yet impair its electricity production. Below we provide an overview of these two major types of PV-system anomalies, and highlight their impact on electricity production.

2.3.1. Intrinsic PV-system faults

Typical PV-system faults include component failure, system isolation due to maintenance work, inverter shutdown due to power cuts or variations in grid voltage, and inverter dropout due to maximum power point tracking (MPPT; Firth et al., 2010). Table 2.1 shows a short description of these common PV-system faults, and indicates their impact on electricity production. Most PV-system faults analyzed by Firth et al. (2010) lead to episodes of brief or sustained zero-production, whereas nonzero production occurs only in case of inverter dropout.

Table 2.1 – Intrinsic PV-system faults as proposed by Firth et al. (2010).

Fault	Description	Impact
Component failure	No production due to component failure	Sustained zero-production
Sustained system isolation	No production when system is switched off (e.g., for maintenance work)	Sustained zero-production
Inverter shutdown	No production due to power cut or variation in grid voltage	Brief zero-production
Brief system isolation	No production when system is switched off (e.g., for maintenance work)	Brief zero-production
Inverter dropout	Reduced production due to MPPT	Reduced nonzero production

2.3.2. Extrinsic factors impairing PV-system performance

There are other factors besides PV-system faults, which are external to PV systems and prevent the generation of maximum electrical power. On the one hand, shading, high temperature, suboptimal tilt/orientation, soiling and high humidity all have a negative impact on PV-system performance. On the other hand, high wind speed simultaneously affects PV-system temperature, humidity and dust deposition, and therefore has an ambiguous impact on PV-system performance. Specifically, high wind speed can either increase power generation by reducing PV-module temperature and humidity, or decrease power generation by scattering dust and thereby causing shading. Table 2.2 lists these common factors affecting PV-system performance, and indicates their impact on electricity production.

Table 2.2 – Extrinsic factors impairing electricity production.

Factor	**Impact**	**Source**
Shading	Up to 79% decrease in power generation	Alonso-García et al. (2006)
High temperature	Up to 15% decrease in power generation	Skoplaki & Palyvos (2009)
Suboptimal tilt/orientation	Up to 59% decrease in power generation	Hussein et al. (2004)
Soiling	Up to 90% decrease in power generation	Sayyah et al. (2014)
High humidity	Up to 50% decrease in power generation	Mekhilef et al. (2012)
High wind speed	Increase or decrease in power generation	Mekhilef et al. (2012)

3. METHODOLOGY

3.1. DATA PREPROCESSING

CSide's time-series of electricity production span several months to years, which is exceedingly long to perform anomaly detection on whole time-series. We therefore analyzed the time-series for two shorter periods of five weeks with contrasting weather conditions, and performed anomaly detection on the last week of these two periods. Specifically, we performed anomaly detection on the week of August 1, 2016 to August 7, 2016, which was particularly favorable for PV-system activity, and the week of November 18, 2016 to November 24, 2016, which was particularly adverse for PV-system activity.

We note, however, that several time-series have spurious values or missing values, which are due to communication problems and preclude a proper analysis of the data. We therefore only performed anomaly detection on time-series that do not have spurious values, which were detected as nonzero production levels recorded during nighttime (i.e., between 00:00 and 04:00). Furthermore, we discarded time-series that are empty for either of the whole two five-week periods of analysis.

Hence, anomaly detection was performed on both complete and incomplete time-series, and a measure of accuracy was used to determine the reliability of anomaly detection. In particular, the accuracy of anomaly detection corresponds to the ratio between the number of observations in a given time-series, and the maximum possible number of observations that time-series can have. For the purpose of illustration, figures shown in the Results section correspond to complete time-series, the anomaly detection of which was 100% accurate.

Data preprocessing yielded 407 eligible time-series for the favorable week of August 1, 2016 to August 7, 2016, and 886 eligible time-series for the adverse week of November 18, 2016 to November 24, 2016.

3.2. ALGORITHMS FOR ANOMALY DETECTION

Although our literature review suggests that neural networks are the most suitable method of unsupervised anomaly detection, we concluded that this approach is not the most adequate to fully meet our study objectives. Specifically, we aim not only to detect anomalies in PV-system activity, but also to establish simple rules that identify several types of anomalies. In our view, neural networks are indeed well suited for anomaly detection, but less appropriate for anomaly identification. We therefore developed five algorithms for anomaly detection, which also allow the identification of several types of anomalies (Fig. 3.1). The proposed algorithms are applied to the preprocessed data in a stepwise manner, and described in the following sections.

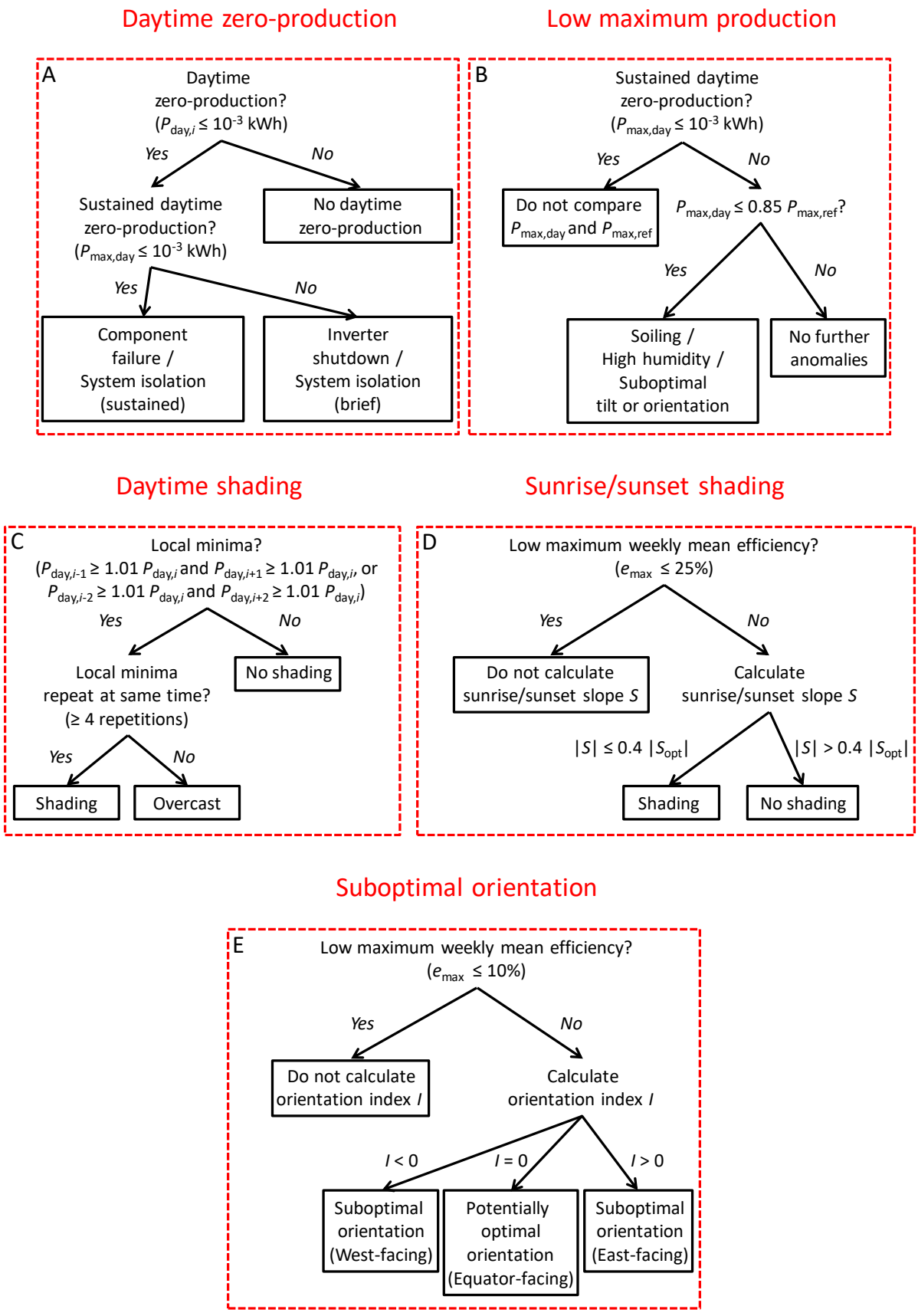

Figure 3.1 – Algorithms used for anomaly detection.

3.2.1. Daytime zero-production

Anomalies leading to daytime zero-production include most of the intrinsic PV-system faults shown in Table 2.1. We assume that time-series with daytime zero-production have at least one daytime observation where power generation, $P_{day,i}$, is sufficiently close to zero (i.e., where $P_{day,i} \leq 10^{-3}$ kWh; see Fig. 3.1*A*).

To determine the daytime period, we first calculate the hour angle, ω_s, at which sunrise ($-\omega_s$) and sunset (ω_s) occur:

$$\omega_S = \begin{cases} 0 & \text{if } -\tan(\phi)\tan(\delta) \geq 1 \\ \pi & \text{if } -\tan(\phi)\tan(\delta) \leq -1, \\ \arccos\left(-\tan(\phi)\tan(\delta)\right) & \text{if } -1 < -\tan(\phi)\tan(\delta) < 1 \end{cases} \quad (7)$$

where ϕ is the latitude at which the PV system is located, and δ is the Sun's declination angle:

$$\delta = 23.45 \frac{\pi}{180} \sin\left(2\pi \frac{284+n}{365.25}\right), \quad (8)$$

which varies daily throughout the year, where n is the day number. Subsequently, we convert the sunrise/sunset hour angle to decimal hours using the Equation of Time (Michalsky, 1988), and define the daytime period as the following interval:

$$\text{day} = [t_{SR} + \text{offset}, t_{SS} - \text{offset}], \quad (9)$$

where t_{SR} and t_{SS} are the sunrise and sunset time in decimal hours, respectively, and offset = 2.5 hours prevents false detections of daytime zero production (e.g., due to sunrise/sunset shading; see below). For simplicity, we assume that the daytime period is the same for all clients analyzed in this study, and perform anomaly detection using the daytime period in Lisbon for August 1, 2016, and November 18, 2016.

If a given time-series has episodes of daytime zero-production, then this algorithm determines whether such episodes are sustained or brief. According to Table 2.1, sustained daytime zero-production can result from component failure or sustained system isolation, whereas brief daytime zero-production can result from inverter shutdown or brief system isolation. On the one hand, sustained daytime zero-production is considered to last at least one day, such that the maximum power generation during daytime in a given day, $P_{max,day}$, is sufficiently close to zero (i.e., $P_{max,day} \leq 10^{-3}$ kWh). On the other hand, brief daytime zero-production is considered to last less than one day, such that the maximum power generation in a given day during daytime is sufficiently higher than zero (i.e., $P_{max,day} > 10^{-3}$ kWh).

3.2.2. Low maximum production

Low maximum production may result from high humidity, soiling, or suboptimal tilt/orientation of the PV system (see Table 2.2). To detect low maximum production, we use an algorithm that determines whether the maximum power generated in a given day, $P_{max,day}$, is sufficiently higher than zero and substantially lower than a reference value, $P_{max,ref}$ (i.e., $10^{-3} < P_{max,day} \leq 0.85\ P_{max,ref}$; see Fig. 3.1*B*). The reference value of maximum power generation corresponds to a standard PV-system

capacity (i.e., 250 W, 500 W, 750 W, 1000 W, and so on), and is calculated based on the historical maximum of power generation. Specifically, the reference value is the closest standard capacity above the historical maximum, where the historical maximum is a median of the 25 highest observations for a given PV system over the analyzed five-week period. For example, if the maximum power generated in a given day is 350 W and the reference value is 500 W, then the PV system is assumed to be operating substantially below capacity (i.e., at 70%) on that day.

3.2.3. Daytime shading

To detect episodes of shading during daytime, we use an algorithm that determines whether a given time-series shows regular local minima (Fig. 3.1*C*). We assume that time-series with local minima have at least one observation that is a local minimum. To ensure the detection of consecutive local minima and avoid the detection of negligible local minima, we adopt a somewhat more stringent definition of local minimum than is common. More specifically, we consider that daytime observation $P_{day,i}$ is a local minimum if the power generation of either both its nearest neighbors, $P_{day,i-1}$ and $P_{day,i+1}$, or both its second-nearest neighbors, $P_{day,i-2}$ and $P_{day,i+2}$, is at least 1% higher than $P_{day,i}$.

If a time-series has local minima, then this algorithm determines whether such local minima occur regularly (i.e., at least 4 days in a week) at a specific time of the day. This repetitive pattern is likely due to shading, which causes a recurrent drop in production at a particular time of the day. Local minima that otherwise occur irregularly are assumed to result from adverse weather conditions (e.g., overcast or rainy weather).

3.2.4. Sunrise/sunset shading

Shading may occur not only during daytime, but also at sunrise/sunset. Shading at sunrise/sunset can be readily detected if the slope of production at sunrise/sunset is substantially less steep than expected. To detect episodes of sunrise/sunset shading, we use an algorithm that compares the steepness of the observed slope of production at sunrise/sunset with that of an optimum slope of production (Fig. 3.1*D*).

To obtain an optimum slope of production at sunrise/sunset, we modeled optimum PV-system efficiency under ideal weather conditions, e_{opt}. Specifically, we first estimated the hourly solar irradiation on a PV-system during August 1, 2016, and November 18, 2016 (Klein, 1977), assuming that the PV system has optimum tilt/orientation and is located in Lisbon. Subsequently, we used the estimated hourly solar irradiation to simulate the efficiency of a model PV-system (Durisch et al., 2007). We refer to the Appendix for a detailed description of our estimation of hourly solar irradiation and optimum PV-system efficiency.

To calculate the observed slope of production at sunrise/sunset, we first draw a curve of weekly mean efficiency of the PV system, e. This curve is obtained by calculating the ratio between the weekly average of electricity production throughout the day, and the reference value of maximum

power generation. Subsequently, we calculate the slope of weekly mean efficiency at sunrise and sunset, S_{SR} and S_{SS}, respectively:

$$S_{SR} = \frac{e_{SR+\text{offset}} - e_{SR}}{\text{offset}}, \tag{10a}$$

$$S_{SS} = \frac{e_{SS} - e_{SS-\text{offset}}}{\text{offset}}, \tag{10b}$$

where e_{SR} and e_{SS} are the weekly mean efficiency at sunrise and sunset, respectively. These observed slopes of weekly mean efficiency are then compared with optimum slopes, $S_{\text{opt},SR}$ and $S_{\text{opt},SS}$, obtained from our PV-system model:

$$S_{\text{opt},SR} = \frac{e_{\text{opt},SR+\text{offset}} - e_{\text{opt},SR}}{\text{offset}}, \tag{11a}$$

$$S_{\text{opt},SS} = \frac{e_{\text{opt},SS} - e_{\text{opt},SS-\text{offset}}}{\text{offset}}, \tag{11b}$$

where $e_{\text{opt},SR}$ and $e_{\text{opt},SS}$ are the optimum efficiency at sunrise and sunset, respectively. Thus, sunrise/sunset shading is detected if observed slopes are at most 40% as steep as optimum slopes (i.e., if $|S_{SR}| \leq 0.4\ |S_{\text{opt},SR}|$ or $|S_{SS}| \leq 0.4\ |S_{\text{opt},SS}|$).

3.2.5. Suboptimal orientation

PV systems with suboptimal orientation are characterized not only by substantial losses in electricity production (see Table 2.2), but also by a temporal mismatch between observed weekly mean efficiency and optimum efficiency (Figs. 3.1*E*, 3.2). We determine the extent to which the orientation of a PV system deviates from optimum conditions, by calculating its orientation index at sunrise and sunset, I_{SR} and I_{SS}, respectively:

$$I_{SR} = t_{\text{opt},SR} - t_{SR}, \tag{12a}$$

$$I_{SS} = t_{\text{opt},SS} - t_{SS}, \tag{12b}$$

where $t_{\text{opt},SR}$ is the moment when optimum efficiency increases to 10% of maximum optimum efficiency, and t_{SR} is the moment when weekly mean efficiency increases to 10% of maximum optimum efficiency. Conversely, $t_{\text{opt},SS}$ is the moment when optimum efficiency drops to 10% of maximum optimum efficiency, and t_{SS} is the moment when weekly mean efficiency drops to 10% of maximum optimum efficiency.

The orientation index of a given client simply corresponds to the average of its sunrise and sunset orientation indices:

$$I = \frac{I_{SR} + I_{SS}}{2}. \tag{13}$$

Thus, the orientation of a PV system is assumed to be Equator-facing and potentially optimal if $I = 0$. Conversely, the orientation of a PV system is assumed to be suboptimal if $I \neq 0$, and considered to be West-facing if $I < 0$ and East-facing if $I > 0$.

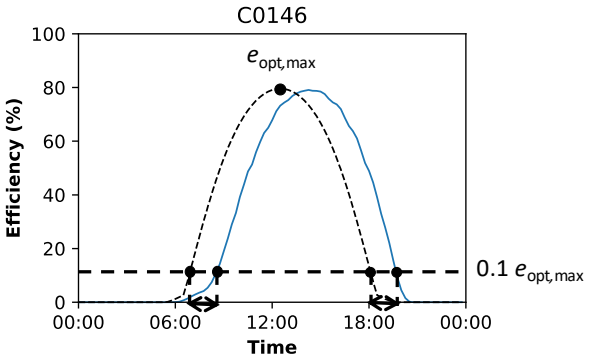

Figure 3.2– Estimation of orientation index from a weekly mean efficiency curve. The blue curve indicates the observed weekly mean efficiency, and the black dashed curve indicates the simulated efficiency under optimum conditions. Black solid circles denote the optimum efficiency maximum, $e_{opt,max}$, and points where observed and optimum efficiency increase or decrease to 10% of the optimum efficiency maximum.

3.3. ESTIMATION OF ANOMALY SEVERITY

Although the detection algorithms developed in this study can be used to identify several types of anomalies, these algorithms do not measure the severity of anomalies. However, anomaly severity is important to determine, because it indicates the extent to which a given anomaly impairs PV-system performance. To address this issue, below we introduce simple metrics that quantify the severity of two anomalies, namely daytime shading and suboptimal orientation.

3.3.1. Daytime shading magnitude and length

In case daytime shading events are detected in a time-series, we measure the severity of daytime shading by calculating its magnitude and length. To this end, we first draw a curve showing the weekly mean efficiency of the PV system, e. Typically, PV systems affected by daytime shading will have weekly mean efficiency curves with one local minimum, e_{min}, and two local maxima, $e_{max,1}$ and $e_{max,2}$ (Fig. 3.3).

Subsequently, we use the weekly mean efficiency curves to determine the magnitude and length of daytime shading. On the one hand, shading magnitude, M, is calculated as the standardized difference between the expected efficiency in the absence of daytime shading, e_{exp}, and the observed local efficiency minimum:

$$M = \frac{e_{\mathrm{exp}} - e_{\mathrm{min}}}{e_{\mathrm{exp}}}, \tag{14}$$

where the expected efficiency is estimated by linear regression through the two local efficiency maxima (Fig. 3.3*A*). On the other hand, shading length, L, is calculated as the time interval between the moment of one of the local efficiency maxima, $t_{max,1}$ or $t_{max,2}$, and the moment when the observed efficiency matches the expected efficiency, t_{exp} (Fig. 3.3*B*):

$$L = \begin{cases} t_{\text{exp}} - t_{\text{max},1} & \text{if } e_{\text{max},2} > e_{\text{max},1} \\ t_{\text{max},2} - t_{\text{exp}} & \text{if } e_{\text{max},2} < e_{\text{max},1} \end{cases}. \quad (15)$$

Thus, we determine the severity of daytime shading by calculating its magnitude and length, such that severe daytime shading events are assumed to have larger magnitude and last for longer periods of time. More specifically, we classify the severity of daytime shading into three different categories: mild shading if $M \leq 15\%$ and $L \leq 1.5$ hours, severe shading if $M \geq 30\%$ and $L \geq 3$ hours and moderate shading otherwise.

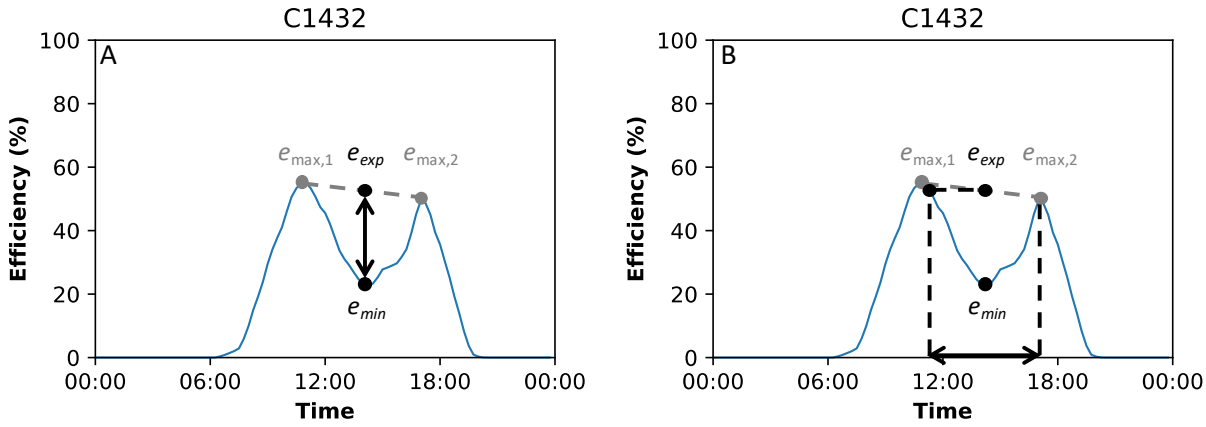

Figure 3.3 – Estimation of daytime shading severity from a weekly mean efficiency curve. *A*, Shading magnitude. *B*, Shading length. Blue curves indicate the observed weekly mean efficiency. Black solid circles denote the local efficiency minimum, e_{min}, and the expected efficiency in the absence of daytime shading, e_{exp}. Gray solid circles denote the two local efficiency maxima, $e_{\text{max},1}$ and $e_{\text{max},2}$, associated with daytime shading.

3.3.2. Orientation index

Similar to daytime shading, we classify the severity of suboptimal orientation into three different categories, which depend on the orientation index described in section 3.2.5. Hence, a PV system is assumed to have mildly suboptimal orientation if $|I| \leq 1$ hour, moderately suboptimal orientation if 1 hour $< |I| \leq 2$ hours, and severely suboptimal orientation if $|I| > 2$ hours.

3.4. ALGORITHM ROBUSTNESS INDICATORS

To evaluate the robustness of our anomaly detection algorithms, we manually annotated time-series with sustained or brief daytime zero-production, daytime shading and sunrise/sunset shading. Subsequently, we used two indicators to quantify the robustness of our anomaly detection algorithms for these anomalies. First, we calculated the detection rate, *DR*, simply as the ratio between the number of time-series correctly detected as anomalous by a given algorithm, *Correct*, and the total number of time-series annotated with a given anomaly, *Annotated*:

$$DR = \frac{Correct}{Annotated} \times 100. \quad (16)$$

Second, we calculated the percentage of false positives, *FP*, as the ratio between the number of time-series incorrectly detected as anomalous by a given algorithm, *Incorrect*, and the total number of time-series detected as anomalous by that algorithm, *Total*:

$$FP = \frac{Incorrect}{Total} \times 100, \quad (17)$$

where *Incorrect* = *Total* – *Correct.* In other words, *FP* is the proportion of time-series identified by a given detection algorithm as anomalous, which are in fact not annotated as such. We therefore assume that the robustness of a given detection algorithm depends on its detection rate (also known as recall or sensitivity) and percentage of false positives, such that robust algorithms will have high detection rate and produce a low percentage of false positives.

We note that time-series annotation is particularly difficult for low maximum production and suboptimal orientation, because we do not have any prior knowledge about the maximum capacity and orientation of the PV systems of our clients. Thus, we did not determine the detection rate and percentage of false positive detections of these two algorithms.

4. RESULTS AND DISCUSSION

4.1. ALGORITHM PERFORMANCE UNDER FAVORABLE WEATHER CONDITIONS

4.1.1. Anomaly detection

The algorithms developed in this study readily detected several types of anomalies under favorable weather conditions (Fig. 4.1). First, 31 clients were detected with sustained daytime zero-production anomalies (Fig. 4.1*A*), and 21 clients with brief daytime zero-production anomalies (Fig. 4.1*B*). Following our time-series annotation, we identified 27 clients with sustained daytime zero-production and 31 clients with brief daytime zero-production. The detection rate of our algorithm for sustained and brief daytime zero-production therefore was 96% (16% false positives) and 61% (9.5% false positives), respectively.

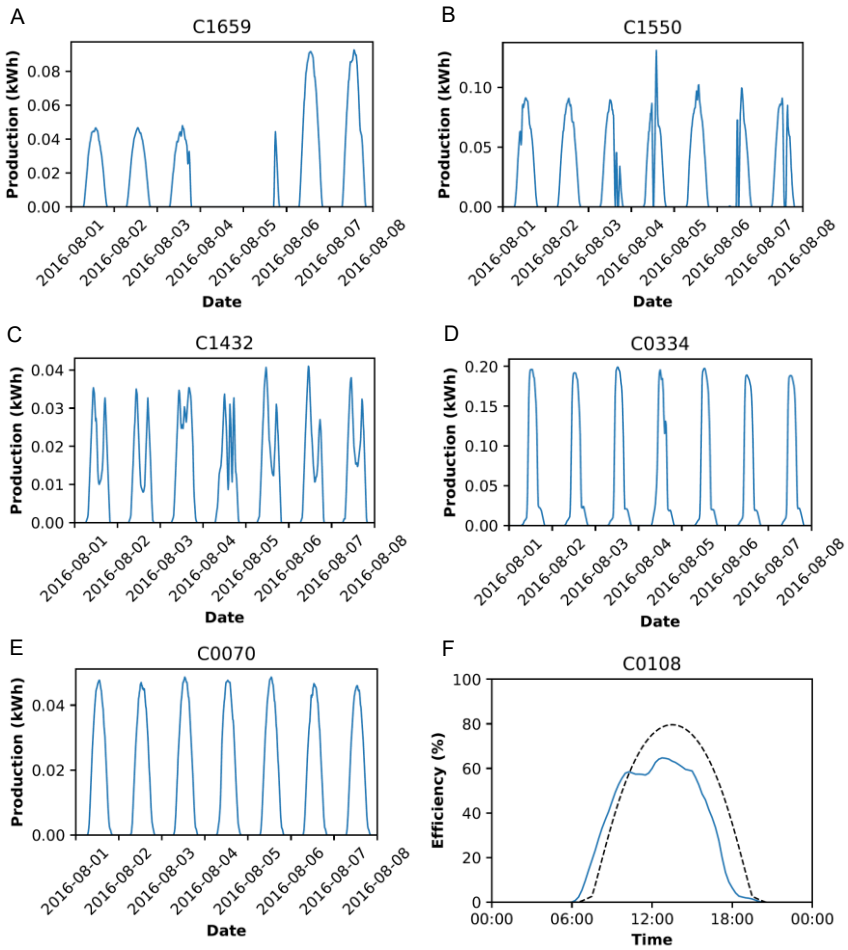

Figure 4.1 – Examples of clients detected with production anomalies under favorable weather conditions. *A,* Client with sustained daytime zero-production. *B,* Client with brief daytime zero-production. *C*, Client with daytime shading. *D*, Client with sunrise and sunset shading. *E*, Client with low maximum production. *F*, Client with suboptimal orientation.

Second, our daytime shading algorithm detected 11 clients with regular local minima (Fig. 4.1*C*), whereas our time-series annotation identified 17 clients with daytime shading. As a result, the

detection rate of our daytime shading algorithm was 65%, and this algorithm did not produce any false positives. Moreover, our sunrise/sunset shading algorithm detected 35 clients with sunrise shading and 98 clients with sunset shading, whereas our time-series annotation identified 53 clients with sunrise shading and 71 clients with sunset shading (Fig. 4.1*D*). Thus, the detection rate of our algorithm for sunrise and sunset shading was 57% (14% false positives) and 96% (31% false positives), respectively.

Third, our algorithms for low maximum production and suboptimal orientation detected 263 clients and 333 clients, respectively, with each type of anomaly (Fig. 4.1*E*, 4.1*F*).

4.1.2. Anomaly severity

4.1.2.1. Daytime shading

To investigate daytime shading severity, we calculated the shading magnitude and length of the 11 clients correctly detected by our algorithm (Table 4.1). Shading magnitude varies considerably among clients, and ranges from M = 7.9% to M = 59.8%. In other words, the local minima of weekly mean efficiency of our clients caused drops between 7.9% and 59.8% relative to expected efficiency in the absence of shading. Similarly, shading length also varies considerably among clients, and ranges from L = 0.75 hours to L = 5.75 hours.

Table 4.1 – Clients correctly detected with daytime shading under favorable weather conditions, and their respective shading magnitude, shading length and shading severity.

Client ID	Shading magnitude, *M* (%)	Shading length, *L* (hours)	Shading severity (-)
C0449	33.5	4.75	Severe
C0494	7.9	1.25	Mild
C0527	13.5	4.25	Moderate
C0536	22.0	1.50	Moderate
C0541	26.2	2.00	Moderate
C0582	59.8	2.00	Moderate
C0773	23.8	4.75	Moderate
C0832	56.9	1.50	Moderate
C0940	14.5	0.75	Mild
C1420	25.3	4.25	Moderate
C1432	55.0	5.75	Severe

The shading magnitude and length of each client should be considered together, so that shading severity can be assessed. Thus, clients mildly affected by daytime shading will have low shading magnitude and short shading length, whereas clients severely affected by daytime shading will have high shading magnitude and long shading length. Figure 4.2 shows the contrast between electricity production and weekly mean efficiency of a client with mild daytime shading (Fig. 4.2*A*, 4.2*C*), and those of a client with severe daytime shading (Fig. 4.2*B*, 4.2*D*).

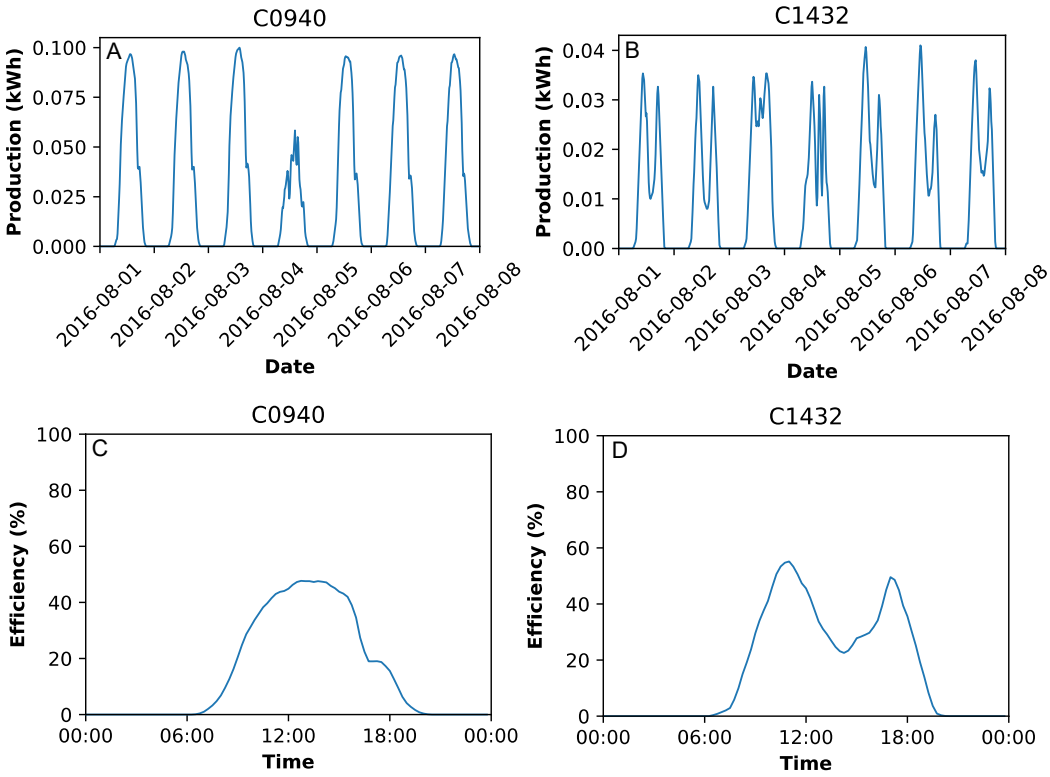

Figure 4.2 – Two clients with contrasting daytime shading severity. *A*, *B*, Time-series of electricity production for (*A*) a client with mild daytime shading (*M* = 14.5%, *L* = 0.75 hours), and (*B*) a client with severe daytime shading (*M* = 55%, *L* = 5.75 hours). *C*, *D*, Weekly mean efficiency curves for (*C*) a client with mild daytime shading, and (*D*) a client with severe daytime shading.

4.1.2.2. Suboptimal orientation

To investigate whether a given PV system has proper orientation, we calculate its orientation index (Fig. 4.3). Similar to daytime shading magnitude and length, the orientation index varies considerably among PV systems, and ranges from $I = -5$ hours to $I = 2$ hours. That is, the orientation index varies between negative and positive values, indicating that the weekly mean efficiency of PV systems with suboptimal orientation (i.e., with $I \neq 0$ hours) can be either lagging or leading relative to the optimum efficiency curve.

From the 353 PV systems eligible for orientation analysis, 20 PV systems have an orientation index of $I = 0$ hours (Fig. 4.3*A*), which indicates a potentially optimal orientation towards the Equator.

Therefore, the majority of PV systems appear to have suboptimal orientation. Specifically, 131 PV systems have negative orientation index and are potentially West-facing (Fig. 4.3*B*), whereas 202 PV systems have positive orientation index and are potentially East-facing (Fig. 4.3*C*).

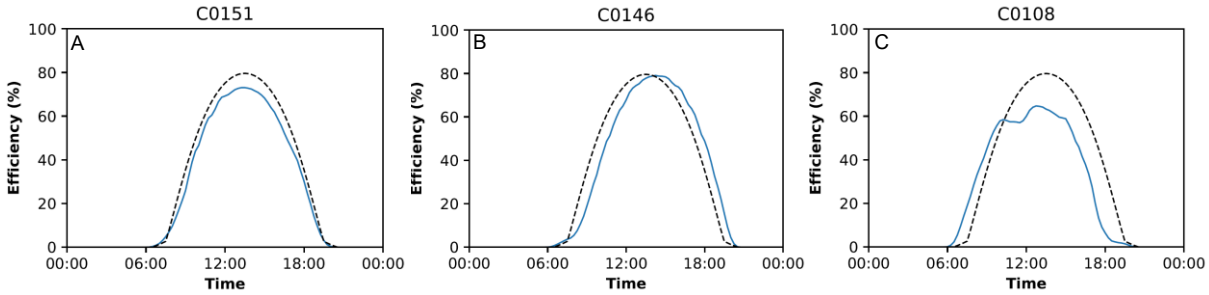

Figure 4.3 – Weekly mean efficiency curves of three PV systems with contrasting orientation index. *A*, PV system with an orientation index of I = 0 hours. *B*, PV system with negative orientation index (I = –0.625 hours). *C*, PV system with positive orientation index (I = 1.125 hours). Blue curves indicate the observed weekly mean efficiency, and black dashed curves indicate the simulated efficiency under optimum conditions.

4.2. ALGORITHM PERFORMANCE UNDER ADVERSE WEATHER CONDITIONS

4.2.1. Anomaly detection

Similar to our anomaly detection under favorable weather conditions, the algorithms developed in this study also detected several types of anomalies under adverse weather conditions (Fig. 4.4). First, 58 clients were detected with sustained daytime zero-production anomalies (Fig. 4.4*A*), and 216 clients with brief daytime zero-production anomalies (Fig. 4.4*B*). Following our time-series annotation, we identified 49 clients with sustained daytime zero-production and 143 clients with brief daytime zero-production. The detection rate of our algorithm for sustained and brief daytime zero-production therefore was 100% (16% false positives) and 67% (56% false positives), respectively.

Second, our daytime shading algorithm detected 61 clients with regular local minima (Fig. 4.4*C*), whereas our time-series annotation identified 26 clients with daytime shading. As a result, the detection rate of our daytime shading algorithm was 69% (71% false positives). Moreover, our sunrise/sunset shading algorithm detected 589 clients with sunrise shading and 373 clients with sunset shading, whereas our time-series annotation identified 415 clients with sunrise shading and 134 clients with sunset shading (Fig. 4.4*D*). Thus, the detection rate of our algorithm for sunrise and sunset shading was 91% (35% false positives) and 78% (72% false positives), respectively.

Third, our algorithms for low maximum production and suboptimal orientation detected 643 clients and 796 clients, respectively, with each type of anomaly (Fig. 4.4*E*, 4.4*F*).

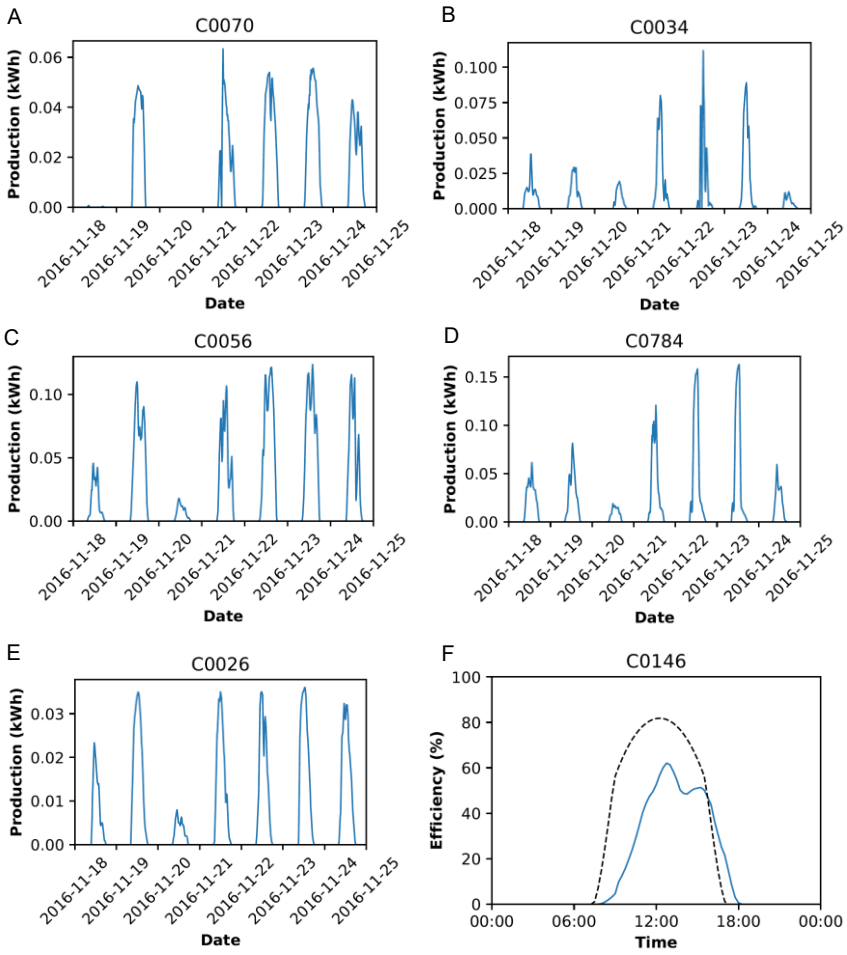

Figure 4.4 – Examples of clients detected with production anomalies under adverse weather conditions. *A,* Client with sustained daytime zero-production. *B,* Client with brief daytime zero-production. *C*, Client with daytime shading. *D*, Client with sunrise and sunset shading. *E*, Client with low maximum production. *F*, Client with suboptimal orientation.

4.2.2. Anomaly severity

4.2.2.1. Daytime shading

We calculated the shading magnitude and length of the 16 clients correctly detected by our algorithm (Table 4.2). Similar to our analysis under favorable weather conditions, shading magnitude varies considerably among clients and ranges from M = 0.6% to M = 60.7%. In other words, the local minima of weekly mean efficiency of our clients caused drops between 0.6% and 60.7% relative to expected efficiency in the absence of shading. Similarly, shading length also varies considerably among clients, and ranges from L = 0.75 hours to L = 3.25 hours.

Table 4.2 – Clients correctly detected with daytime shading under adverse weather conditions, and their respective shading magnitude, shading length and shading severity.

Client ID	Shading magnitude, *M* (%)	Shading length, *L* (hours)	Shading severity (-)
C0056	21.1	1.50	Moderate
C0123	35.7	1.75	Moderate
C0144	5.4	0.75	Mild
C0148	19.5	1.50	Moderate
C0211	60.7	1.50	Moderate
C0221	24.1	1.25	Moderate
C0323	13.5	1.50	Mild
C0467	44.9	1.25	Moderate
C0491	7.2	1.00	Mild
C0689	13.2	0.75	Mild
C0870	0.6	3.25	Moderate
C1091	19.0	3.25	Moderate
C1196	25.6	3.25	Moderate
C1338	44.9	1.50	Moderate
C1643	34.9	1.50	Moderate
C1648	4.6	3.25	Moderate

4.2.2.2. Suboptimal orientation

The orientation index varies considerably among PV systems under adverse weather conditions, and ranges from $I = -4$ hours to $I = 1.125$ hours. From the 813 PV systems eligible for orientation analysis, 17 PV systems have an orientation index of $I = 0$ hours, which indicates a potentially optimal orientation towards the Equator. Therefore, the majority of PV systems appear to have suboptimal orientation. Specifically, 784 PV systems have negative orientation index and are potentially West-facing, whereas 12 PV systems have positive orientation index and are potentially East-facing.

4.3. DISCUSSION

4.3.1. Anomaly detection

Our results indicate that the algorithms proposed in this study perform quite well on anomaly detection under favorable weather conditions. In particular, our algorithms successfully detected the majority of clients labeled with either sustained or brief daytime zero-production, and either daytime or sunrise/sunset shading (Table 4.3). Importantly, these algorithms also produced a relatively low percentage of false positives, which indicates that most anomaly detections are correct.

The detection rate of our algorithms is similarly high under adverse weather conditions, and in many cases higher than under favorable weather conditions. However, the percentage of false positive anomaly detections is also substantially higher under adverse weather conditions, which indicates that the algorithms are generally more robust under favorable weather conditions.

Table 4.3 – Algorithm performance under favorable versus adverse weather conditions.

	August 1-7, 2016 (favorable conditions)		November 18-24, 2016 (adverse conditions)	
Anomaly	**Detection rate (%)**	**False positives (%)**	**Detection rate (%)**	**False positives (%)**
Sustained zero-production	96	16	100	16
Brief zero-production	61	9.5	67	56
Daytime shading	65	0	69	71
Sunrise shading	57	14	91	35
Sunset shading	96	31	78	72

Our analysis also suggests that a majority of PV systems either has low maximum production, or is suboptimally oriented. On the one hand, 263 PV systems (i.e., 65% of the analyzed clients) were detected with low maximum production, whereas 333 PV systems (i.e., 82% of the analyzed clients) were detected with suboptimal orientation under favorable weather conditions. On the other hand, 643 PV systems (i.e., 73% of the analyzed clients) were detected with low maximum production, whereas 796 PV systems (i.e., 90% of the analyzed clients) were detected with suboptimal orientation under adverse weather conditions. These results indicate that low maximum production and suboptimal orientation are particularly prevalent, and that our detection algorithms may be useful to alert clients for these two common types of anomalies.

4.3.2. Anomaly severity

An important goal of this study is to determine anomaly severity, which measures the impact of anomalies on PV-system activity. Our metrics for anomaly severity seem to perform well, and

indicate that daytime shading and suboptimal orientation can have a substantial impact on PV-system activity (Figs. 4.2, 4.3). On the one hand, daytime shading can lead to efficiency losses of up to 60% (see Tables 4.1, 4.2), which are in agreement with a decrease in power generation of up to 79% reported by Alonso-García et al. (2006). Although we only detected daytime shading in less than 5% of our clients, this type of anomaly can have a strong impact on PV-system performance and should therefore be taken into account.

On the other hand, suboptimal orientation can shift the weekly mean efficiency curve of a PV system, and thereby lead to large mismatches relative to the optimum efficiency curve (Fig. 4.3). Indeed, suboptimal orientation may cause the PV system to either lag up to 5 hours behind the optimum efficiency curve, or lead up to 2 hours ahead of the optimum efficiency curve. Although we did not measure losses in PV-system efficiency associated with such mismatches, theoretical studies show that suboptimal orientation can drive a substantial decrease in power generation (Hussein et al., 2004). Given that suboptimal orientation affects more than 80% of our clients, the impact of this prevalent anomaly on PV-system efficiency merits further investigation.

5. CONCLUSIONS

We develop five algorithms for the detection of several PV-system anomalies, and establish metrics to determine the severity of daytime shading and suboptimal orientation. Specifically, our algorithms are used to detect brief and sustained daytime zero-production, daytime and sunrise/sunset shading, low maximum production and suboptimal orientation. We apply these detection algorithms to several time-series of electricity production, which were obtained for two periods with contrasting weather conditions. When weather conditions were favorable, our algorithms successfully detected the majority of time-series labeled with either sustained or brief daytime zero-production, and either daytime or sunrise/sunset shading. Furthermore, these algorithms also produced a relatively low percentage of false positives, which indicates that most anomaly detections are correct. When weather conditions were adverse, the detection rate of our algorithms was similarly high, if not higher, than when weather conditions were favorable. However, the percentage of false positive anomaly detections is also substantially higher under adverse weather conditions, which indicates that the algorithms are generally more robust under favorable weather conditions.

Shading severity varied substantially among PV systems, and caused efficiency losses of up to 60.7%. Such large efficiency losses indicate that shading can have a strong impact on PV-system performance and should therefore be taken into account, even though it only affects a small proportion of PV systems. Suboptimal orientation was detected on more than 80% of the PV systems analyzed in this study, and drove a temporal shift in observed PV-system efficiency relative to optimum efficiency. The high prevalence of suboptimal orientation suggests that this type of anomaly is rather common, and therefore warrants further investigation into its impact on PV-system efficiency.

5.1. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Our results suggest that the approach developed in this study is well capable of detecting several types of PV-system anomalies and estimating their severity, especially when weather conditions are favorable. Yet, our study has two important limitations, which deserve further scrutiny and should be addressed in future work. First, we analyzed time-series of electricity production for the weeks of August 1, 2016 to August 7, 2016, and November 18, 2016 to November 24, 2016, which had contrasting weather conditions for PV-system activity. To better assess the robustness of this framework, however, our detection algorithms and severity metrics should be tested on more weeks of PV-system activity. In particular, our framework should be applied to at least one week of each season of the year, in order to accommodate seasonal changes in electricity production. Only then we may draw more definite conclusions about the robustness of our detection algorithms and severity metrics.

Second, most of our algorithms perform anomaly detection during the daytime period, which is defined as the interval between sunrise and sunset minus an offset of 2.5 hours (see Equation [9]). That is, the offset used in our algorithms effectively reduces the daytime period by 5 hours. Although this offset decreases the number of false positive detections, it also reduces the detection rate of our algorithms. This problem is particularly pertinent if the analyzed PV systems are located at high

latitudes during wintertime, in which case this daytime period becomes too short for anomaly detection. Thus, a more realistic definition of daytime period will be necessary for our algorithms to detect anomalies on any PV system, regardless of its location and time of the year.

REFERENCES

Alonso-García, M. C., Ruiz, J. M., & Chenlo, F. (2006). Experimental study of mismatch and shading effects in the *I–V* characteristic of a photovoltaic module. *Solar Energy Materials and Solar Cells*, 90, 329–340.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons, 3rd edition.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: a review. *Statistical Science*, 17(3), 235–249.

BP. (2018). Solar energy. Retrieved September 20, 2018, from https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy/renewable-energy/solar-energy.html.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. *ACM Computing Surveys*, 41(3), 15:1–15:58.

Chouder, A., & Silvestre, S. (2010). Automatic supervision and fault detection of PV systems based on power losses analysis. *Energy Conversion and Management*, 15, 1929–1937.

Dimroth, F., Grave, M., Beutel, P., Fiedeler, U., Karcher, C., Tibbits, T. N. D., Oliva, E., Siefer, G., Schachtner, M., Wekkeli, A., Bett, A. W., Krause, R., Piccin, M., Blanc, N., Drazek, C., Guiot, E., Ghyselen, B., Salvetat, T., Tauzin, A., Signamarcheix, T., Dobrich, A., Hannappel, T., & Schwarzburg, T. (2014). Waferbonded four-junction GaInP/GaAs//GaInAsP/GaInAs concentrator solar cells with 44.7% efficiency. *Progress in Photovoltaics: Research and Applications*, 22, 277–282.

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, 548–564.

Durisch, W., Bitnar, B., Mayor, J. C., Kiess, H., Lam, K., & Close, J. (2007). Efficiency model for photovoltaic modules and demonstration of its application to energy yield estimation. *Solar Energy Materials & Solar Cells*, 91, 79–84.

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.

Gopinathan, K. K. (1991). Solar radiation on variously oriented sloping surfaces. *Solar Energy*, 47, 173–179.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1),27–58.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85–126.

Houghton, J. T., Callander, B. A., & Varney, S. K. (Eds.). (1992). *Climate change 1992*. Cambridge University Press.

Hüsken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing*, 50, 223–235.

Hussein, H. M. S., Ahmad, G. E., & El-Ghetany, H. H. (2004). Performance evaluation of photovoltaic modules at different tilt angles and orientations. *Energy Conversion and Management*, 45, 2441–2452.

IEA (2014a). *Technology roadmap: solar photovoltaic energy*. International Energy Agency, Paris.

IEA (2014b). *Energy technology perspectives 2014*. International Energy Agency, Paris.

Klein, S. A. (1977). Calculation of monthly average insolation on tilted surfaces. *Solar Energy*, 19, 325–329.

Knorr, M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *International Conference on Very Large Data Bases* (pp. 392-403).

Kopp, G., & Lean, J. L. (2011). A new, lower value of total solar irradiance: evidence and climate significance. *Geophysical Research Letters*, 38, L01706.

Lashof, D. A., & Ahuja, D. R. (1990). Relative contributions of greenhouse gas emissions to global warming. *Nature*, 344(6266), 529–531.

Laurikkala, J., Juhola, M., & Kentala, E. (2000). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology* (Vol. 1, pp. 20–24).

Långkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42, 11–24.

Madsen, J. H. (2018). *Outlier detection for improved clustering* (M.Sc. thesis). NOVA Information Management School.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv*, abs/1607.00148.

Markou, M., & Singh, S. (2003a). Novelty detection: a review — part 1: statistical approaches. *Signal Processing*, 83, 2481–2497.

Markou, M., & Singh, S. (2003b). Novelty detection: a review — part 2: neural network based approaches. *Signal Processing*, 83, 2499–2521.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., Frame, D. J., & Allen, M. R. (2009). Greenhouse-gas emission targets for limiting global warming to 2°C. *Nature*, 458(7242), 1158–1163.

Mekhilef, S., Saidur, R., & Kamalisarvestani, M. (2012). Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. *Renewable and Sustainable Energy Reviews*, 16, 2920–2925.

Michalsky, J. J. (1988). The Astronomical Almanac's algorithm for approximate solar position (1950-2050). *Solar Energy*, 40, 227–235.

Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*. URL: https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf.

Pereira, J. P. C. (2018). *Autoencoders for unsupervised anomaly detection in time series data* (M.Sc. report). Instituto Superior Técnico.

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.

Rahman, H., Pinty, B., & Verstraete, M. M. (1993). Coupled surface-atmosphere reflectance (CSAR) model 2. Semiempirical surface model usable with NOAA advanced very high resolution radiometer data. *Journal of Geophysical Research*, 98, 20791–20801.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 427–438). ACM Press.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 73–79.

Sayyah, A., Horenstein, M. N., & Mazumder, M. K. (2014). Energy yield loss caused by dust deposition on photovoltaic panels. *Solar Energy*, 107, 576–604.

Skoplaki, E., & Palyvos, J. A. (2009). On the temperature dependence of photovoltaic module electrical performance: a review of efficiency/power correlations. *Solar Energy*, 83, 614–624.

Solomon, S., et al. (2007). Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, 2007.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

UNFCCC. (2018). The Paris agreement. United Nations Framework Convention on Climate Change. Retrieved September 20, 2018, from https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1), 1–17.

Woyte, A., Belmans, R., & Nijs, J. (2007). Fluctuations in instantaneous clearness index: analysis and statistics. *Solar Energy*, 81, 195–206.

6. APPENDIX

6.1. HOURLY SOLAR IRRADIATION ON A TILTED PV-SYSTEM

To detect sunrise/sunset shading and suboptimal orientation, we use optimum PV-system efficiency curves that are obtained from estimations of hourly solar irradiation on a tilted PV system. The hourly solar irradiation on a tilted PV system, $\dot{H}(\beta)$, is estimated as follows (Klein, 1977):

$$\dot{H}(\beta) = \dot{B}\dot{R}_b + \dot{D}\left(\frac{1+\cos(\beta)}{2}\right) + \left(\frac{1-\cos(\beta)}{2}\right)\rho\dot{H}, \quad \text{(A1)}$$

where $\dot{B}$ is the hourly beam component of solar irradiation, $\dot{R}_b$ is the ratio of the average beam radiation on a tilted surface to that on a horizontal surface, $\dot{D}$ is the hourly diffused irradiation for the average day of each month, β is the tilt of the PV system from horizontal, ρ is the ground reflectance and $\dot{H}$ is the total hourly irradiation for the average day of each month. Below we briefly explain how the terms in Equation (A1) are obtained.

The total hourly irradiation for the average day of each month can be estimated as follows:

$$\dot{H} = r_t\overline{H}, \quad \text{(A2)}$$

where r_t and $\overline{H}$ are the average hourly irradiation for each month and the monthly average daily irradiation on a horizontal plane, respectively. The monthly average daily irradiation on a horizontal plane corresponds to the amount of extraterrestrial irradiation that reaches Earth's surface:

$$\overline{H} = \overline{K}_T\overline{H}_{0h}, \quad \text{(A3)}$$

where $\overline{K}_T$ is the clearness index, and $\overline{H}_{0h}$ is the total amount of extraterrestrial irradiation reaching Earth's atmosphere. The total amount of extraterrestrial irradiation reaching Earth's atmosphere is a fraction of the solar constant, I_{SC}, and varies throughout the year:

$$\overline{H}_{0h} = \frac{24}{\pi}I_{SC}\left[1 + 0.034\cos\left(2\pi\frac{n}{365.25}\right)\right]\left(\cos(\phi)\cos(\delta)\sin(\omega_s) + \omega_s\sin(\phi)\sin(\delta)\right), \quad \text{(A4)}$$

where n is the day number, ϕ is the latitude at which the PV system is located, δ is the Sun's declination angle and ω_s is the hour angle at which sunrise and sunset occur:

$$\omega_s = \begin{cases} 0 & \text{if } -\tan(\phi)\tan(\delta) \geq 1 \\ \pi & \text{if } -\tan(\phi)\tan(\delta) \leq -1 \\ \arccos\left(-\tan(\phi)\tan(\delta)\right) & \text{if } -1 < -\tan(\phi)\tan(\delta) < 1 \end{cases}. \quad \text{(A5)}$$

Thus, at a given latitude, the extraterrestrial irradiation reaching Earth's atmosphere is essentially driven by the Sun's declination angle:

$$\delta = 23.45\frac{\pi}{180}\sin\left(2\pi\frac{284+n}{365.25}\right). \quad \text{(A6)}$$

The average hourly irradiation for each month varies throughout the day, and is calculated as follows:

$$r_t = \begin{cases} \frac{\pi}{24} r_0 & \text{if } r_0 > 0 \\ 0 & \text{if } r_0 \leq 0 \end{cases}, \tag{A7}$$

where $r_0 = \frac{\cos(\omega) - \cos(\omega_s)}{\sin(\omega_s) - \omega_s \cos(\omega_s)}$ and ω is the hour angle. The hourly beam component of solar irradiation corresponds to the total hourly irradiation that is not diffused:

$$\dot{B} = \dot{H} - \dot{D}. \tag{A8}$$

The hourly diffused irradiation for the average day of each month:

$$\dot{D} = r_d \overline{D}, \tag{A9}$$

depends on the diffused component of total irradiation:

$$r_d = \begin{cases} \frac{\pi}{24} r_0 & \text{if } r_0 > 0 \\ 0 & \text{if } r_0 \leq 0 \end{cases}, \tag{A10}$$

and on the monthly average daily diffused irradiation on a horizontal surface:

$$\overline{D} = \begin{cases} (1.391 - 3.560\overline{K}_T + 4.189\overline{K}_T^2 - 2.137\overline{K}_T^3)\overline{H} & \text{if } \omega_s < 81.4° \\ (1.311 - 3.022\overline{K}_T + 3.427\overline{K}_T^2 - 1.821\overline{K}_T^3)\overline{H} & \text{if } \omega_s \geq 81.4° \end{cases}. \tag{A11}$$

If we assume that the PV system is optimally oriented towards the equator, then the ratio of the average beam radiation on a tilted surface to that on a horizontal surface can be calculated as follows:

$$\dot{R}_b = \begin{cases} 0 & \text{if } \frac{\cos(\theta_i)}{\cos(\theta_Z)} \leq 0 \\ \frac{\cos(\theta_i)}{0.25} & \text{if } 0 \leq \cos(\theta_Z) < 0.25 \\ \frac{\cos(\theta_i)}{-0.25} & \text{if } -0.25 < \cos(\theta_Z) \leq 0 \\ \frac{\cos(\theta_i)}{\cos(\theta_Z)} & \text{otherwise} \end{cases}, \tag{A12}$$

where

$$\frac{\cos(\theta_i)}{\cos(\theta_Z)} = \frac{\cos(\delta)\cos(\phi-\beta)\cos(\omega)+\sin(\delta)\sin(\phi-\beta)}{\cos(\delta)\cos(\phi)\cos(\omega)+\sin(\delta)\sin(\phi)}. \tag{A13}$$

We refer to Table 6.1 for the parameters used to estimate hourly solar irradiation on a tilted PV-system.

Table 6.1 – Parameters used to estimate hourly solar irradiation on a tilted PV-system.

Symbol	Definition	Unit	Value	Source
I_{SC}	Solar constant	W m^{-2}	1367	Kopp & Lean (2011)
ρ	Ground reflectance	-	0.1	Rahman et al. (1993)
$\overline{K}_T$	Clearness index	-	0.75	Woyte et al. (2007)
ϕ	Latitude	radians	40π/180	-
β	PV-system tilt	radians	ϕ	Gopinathan (1991)

6.2. PV-SYSTEM MODEL

The hourly solar irradiation on a tilted PV-system estimated with Equation (A1), $\dot{H}(\beta)$, is used to simulate optimum PV-system efficiency under ideal weather conditions. To this end, we first calculate the efficiency of a solar cell following Durisch et al. (2007):

$$\eta = \frac{p}{100}\left[q\frac{\dot{H}(\beta)}{\dot{H}_0(\beta)} + \left(\frac{\dot{H}(\beta)}{\dot{H}_0(\beta)}\right)^m\right]\left(1 + r\frac{T_c}{T_0}\right), \quad \text{(A14)}$$

where p, q, r and m are regression parameters empirically estimated by Durisch et al. (2007), $\dot{H}_0(\beta)$ and T_0 are the hourly solar irradiation and solar cell temperature at standard testing conditions (STC), respectively, and T_c is the solar cell temperature. The solar cell temperature is assumed to depend on air temperature, T_a, and increase with solar irradiation:

$$T_c = T_a + h\dot{H}(\beta), \quad \text{(A15)}$$

where h is the Ross coefficient, which measures the warming rate of the solar cell with solar irradiation. In line with expectation, Equation (A14) predicts that solar cell efficiency will tend to increase with solar irradiation, and decrease with solar cell temperature. Solar cell efficiency can be subsequently used to estimate optimum PV-system efficiency:

$$e_{\text{opt}} = \frac{\eta A\dot{H}(\beta)}{250}, \quad \text{(A16)}$$

where A = 1.6 m^2 is the area of a typical PV system with 250 W of power generation capacity. Model parameters are listed in Table 6.2.

Table 6.2 – PV-system model parameters.

Symbol	Definition	Unit	Value	Source
p	Empirical regression parameter	-	24	Durisch et al. (2007)
q	Empirical regression parameter	-	-0.3	Durisch et al. (2007)
r	Empirical regression parameter	-	-0.1	Durisch et al. (2007)
m	Empirical regression parameter	-	0.2	Durisch et al. (2007)
h	Ross coefficient	°C $(W\ m^{-2})^{-1}$	0.03	Durisch et al. (2007)
T_0	Solar cell temperature (STC)	°C	25	Durisch et al. (2007)
$\dot{H}_0(\beta)$	Hourly solar irradiation (STC)	$W\ m^{-2}$	1000	Durisch et al. (2007)
T_a	Air temperature	°C	15–25	-
A	PV-system area	m^2	1.6	-