energies

MDPI

Article

Anomaly Detection in Photovoltaic Production Factories via Monte Carlo Pre-Processed Principal Component Analysis

Eleonora Arena [1], Alessandro Corsini [2], Roberto Ferulano [3], Dario Alfio Iuvara [1], Eric Stefan Miele [2], Lorenzo Ricciardi Celsi [3,*], Nour Alhuda Sulieman [4] and Massimo Villari [4]

1 Enel Green Power S.p.A., Contrada Blocco Torrazze sn, Zona Industriale, 95121 Catania, Italy; eleonora.arena@enel.com (E.A.); dario.iuvara@enel.com (D.A.I.)
2 Dipartimento di Ingegneria Astronautica, Elettrica ed Energetica, Sapienza Università di Roma via Eudossiana 18, 00184 Roma, Italy; alessandro.corsini@uniroma1.it (A.C.); ericstefan.miele@uniroma1.it (E.S.M.)
3 ELIS Innovation Hub, via Sandro Sandri 81, 00159 Roma, Italy; r.ferulano@elis.org
4 Dipartimento di Scienze Matematiche e Informatiche, Scienze Fisiche e Scienze Della Terra, Università di Messina, Piazza Pugliatti 1, 98122 Messina, Italy; nosulieman@unime.it (N.A.S.); mvillari@unime.it (M.V.)
* Correspondence: l.ricciardicelsi@elis.org

Abstract: This paper investigates a use case of robust anomaly detection applied to the scenario of a photovoltaic production factory—namely, Enel Green Power's 3SUN solar cell production plant in Catania, Italy—by considering a Monte Carlo based pre-processing technique as a valid alternative to other typically used methods. In particular, the proposed method exhibits the following advantages: (i) Outlier replacement, by contrast with traditional methods which are limited to outlier detection only, and (ii) the preservation of temporal locality with respect to the training dataset. After pre-processing, the authors trained an anomaly detection model based on principal component analysis and defined a suitable key performance indicator for each sensor in the production line based on the model errors. In this way, by running the algorithm on unseen data streams, it is possible to isolate anomalous conditions by monitoring the above-mentioned indicators and virtually trigger an alarm when exceeding a reference threshold. The proposed approach was tested on both standard operating conditions and an anomalous scenario. With respect to the considered use case, it successfully anticipated a fault in the equipment with an advance of almost two weeks, but also demonstrated its robustness to false alarms during normal conditions.

Keywords: anomaly detection; principal component analysis; Monte Carlo simulation; PV cell production line; predictive maintenance

Citation: Arena, E.; Corsini, A.; Ferulano, R.; Iuvara, D.A; Miele, E.S; Ricciardi Celsi, L.; Sulieman, N.A; Villari, M. Anomaly Detection in Photovoltaic Production Factories via Monte Carlo Pre-Processed Principal Component Analysis. *Energies* **2021**, *14*, 3951. https://doi.org/10.3390/en14133951

Academic Editor: Lyes Bennamoun

Received: 14 May 2021
Accepted: 25 June 2021
Published: 1 July 2021

1. Introduction

In recent years, predictive maintenance has been receiving an ever increasing attention and has been considered fundamental in industrial applications. In fact, it contributes to guaranteeing healthy, safe and reliable systems, as well as to avoiding breakdowns that could potentially lead to a whole system shutdown.

As known, the main benefit of Principal Component Analysis (PCA) lies in its capability to reduce the dimensionality of data by selecting the most important features that are responsible for the highest variability in the input dataset. Namely, PCA allows to concentrate the analysis on a compressed version of the original dataset without compromising the reliability and the robustness of a predictive model. Among other factors, a key quality in PCA is the inherent capability of processing large multivariate datasets as customary in industrial equipment sensor networks. As a result, PCA formed a field of choice in predictive analytics in several use cases, e.g., maritime and transport applications, as well as decision support systems in healthcare [1,2].

On the other hand, the well known disadvantage of PCA stems from the sensitivity to outliers in the data. In this respect, in the literature four known algorithms have been very

recently devised in order to sort outliers' observations out, namely the spherical principal component based algorithm, PCA based on robust covariance matrix estimation, robust PCA (ROBPCA) and the PCA projection pursuit algorithm [3].

To this end, based on measurements collected by the sensor network of a photovoltaic production plant, the paper proposes Monte Carlo (MC) simulation as the pre-processing stage to deal with outliers before applying PCA [4,5]. In this respect, the proposed approach is shown to be a valid alternative to relying on the classical Interquartile Range (IQR) method in order to omit outliers when applying PCA for anomaly detection purposes.

1.1. Related Works

Recently, the scientific community has devoted much attention to the use of data analytics and machine learning models in the operation domains, e.g., manufacturing and energy management. In particular, many applications have focused on predictive maintenance and anomaly detection [6–8].

In this context, industrial systems have adopted PCA for detecting anomalous scenarios in their operational processes. In particular, key performance indicators (KPIs) are usually defined starting from the PCA model in order to trigger alarms and prevent failures [9].

Many works focus on fault isolation techniques which are employed to classify different occurring errors and to isolate the system variables mostly affected by them [10]. Specifically, they often propose statistical methods for fault detection, like Hotelling T^2 or squared prediction errors Q [11,12].

Even though plenty of these works deal with error classification and isolation in the context of anomaly detection and predictive maintenance, other papers and practical experiments shed light on innovative strategies to pre-process the input data that will feed the predictive model. To this end, MC simulation has been largely applied for data pre-processing in order to define more robust models. For example, in [13] the authors process geodetic data by applying MC simulation to perform uncertainty modelling [14].

However, choosing the statistical method for MC simulation becomes difficult when the involved dataset is highly affected by the presence of outliers. In this respect, a robust estimation procedure has been investigated in [15]: The authors exploit the median since it provides an estimator with the highest breakdown point and it always guarantees a feasible solution for the considered optimization problem.

In general, MC simulation is used as a valid pre-processing strategy in order to successfully manage uncertainty with respect to experimental use cases in manufacturing and energy management, namely for predictive maintenance [16–19] or predictive analytics purposes [20].

Moreover, the number of data points sampled by MC simulation is another crucial parameter, since it could lead to inaccurate outputs [21]. This parameter is particularly challenging to optimize since it strongly depends on the use case and the quality of data. In [22] the authors test different MC simulations to determine the relationship between the sample size and the accuracy of the sample mean and variance.

Despite larger samples could provide for a better estimation of the input distributions, in [23] results demonstrated the need to restrict the number of MC runs to a number not greater than the sample sizes used for the input parameters, since a large number could be unnecessary or even harmful.

Despite the clear advantage of such approaches, they often still need to be validated in practice. So, to the best of the authors' knowledge, this paper proposes the application of MC simulation to a real photovoltaic production scenario, as an effective way to pre-process the data stream coming from the sensors deployed throughout the production site.

The related literature also reports pre-processing techniques for similar anomaly detection scenarios based on the IQR method (e.g., [24]), which, however, offers only the property of outlier removal and not the additional benefit of outlier replacement that is consequential to applying MC simulation, as further discussed in Section 3.

1.2. Paper Structure

The paper is structured as follows. Section 2 provides the use case description and problem setting. In Section 3 we explain our contribution in terms of exploiting MC simulation as an innovative approach to data pre-processing with respect to the considered anomaly detection and predictive maintenance application. Later on, in Section 4 we discuss PCA for anomaly detection. Section 5 presents the experimental setup and numerical results. Finally, Section 6 concludes the paper.

2. Problem Setting

Enel Green Power needs to implement, in the production line of sun cells in the 3SUN Factory, an artificial intelligence application capable of predicting faults relative to a piece of process equipment, the so-called Automatic Wet Bench (AWB) machine, for predicting any malfunctioning of the fans that ventilate the different stations within such machine. The data collected on the Manufacturing Execution System (MES) are fed as input to the predictive analytics engine in order to predict faults.

2.1. Use Case

In Figure 1 we show the process steps involved in the cell production. Each process equipment has a specific purpose: Raw wafers enter the first machine in the line, the so-called Wafer Inspection System (WIS), to check the quality of the input wafers; then, they are subject to texturization and cleaning through the AWB equipment; next, the Plasma Enhanced Chemical Vapor Deposition (PeCVD) equipment is used for the deposition of doped and un-doped layer of amorphous Silicon (aSi) on both side of the wafers. Then, the Physical Vapour Deposition equipment (PVD) is used for the sputtering process. Finally, the block formed by the Screen Printer, Tester and Sorter equipment are responsible, respectively, of collecting the electric charge of the cell (fingers) and to let the flow between one cell and the other (Bus Bar) in the assembled modules, testing the electrical I-V measurements of the cells and classifying them depending on their perfomance.

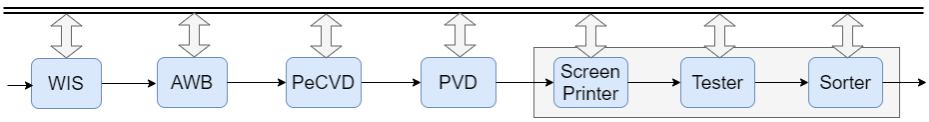


Figure 1. Photovoltaic cell production line in the 3SUN Factory.

The process equipment we refer to in this paper in order to predict the occurrence of faults is the AWB, where the wafers are chemically etched to roughen the surface to maximize the quantity of absorbed light and therefore the cell efficiency.

Along the production line, two parallel AWB machines are installed, each consisting of a loading station (the first one) and an unloading station (the last one) and, midway between the two, several stations where the chemical processes are performed. Within the AWB stage, the wafers are loaded onto specific containers called carriers, which move from one station to another until the process ends; the carriers do not enter in all stations but only some of them, as the same task can be carried out indifferently by one station or another, so that the carrier is moved by the automation system to the first available station that can carry out the required task.

More specifically, the stations composing the production line serve three main purposes: Pre-conditioning, texturing and cleaning. Each station is equipped with a sensor that records measurements when carriers enter and exit the station.

We now provide a brief description of the most frequently occurring fault inside the AWB and for which we design a suitable predictive analytics strategy. Such a fault is generally due to the malfunctioning of the fans that ventilate the different stations within each AWB stage.

For each AWB stage, there exist two drying tanks which must work properly in parallel and can never break down (not even alternatively), otherwise the AWB throughput would be halved, thus compromising the whole production line. Since the fault episode is generally preceded by the occurrence of anomalous vibrations, there is room for a suitable predictive analytics strategy aimed at anticipating the occurrence of the fault through the detection of such vibrations.

At a specific slot of time, an unexpected error may happen in one of its machines and block the production completely for several few days.

2.2. Sensor Measurements

The sensors mounted onto the production line stations measure several relevant parameters characterizing each station, such as station temperature, pump speed, flow speed, and ozone concentration level.

The measurements recorded by the sensors were collected only during the enter, exit and dosing phases of each carrier, thus leading to a non-constant sampling frequency. This produced many discontinuities of variable length in the sensor data streams, making standard time series analysis impossible. For this reason, the collected measurements were treated as an ordered set of samples rather than time series. In order to capture the time evolution of carriers going through a line, each sample is composed by the measurements coming from all the stations, collected during the enter, exit and dosing phases of a carrier.

Let k stations out of the total number N account for the main path drawn by a carrier entering the AWB stage to undergo pre-conditioning, texturing and cleaning. The remaining $(N-k)$ stations are parallel to the k principal ones and ensure the robustness of the whole AWB stage in the following way: If one of the k stations fails, there is at least a redundant station among the available $(N-k)$ that is properly working and can thus be entered by the carrier to undergo the whole production process.

For the sake of simplicity and without loss of generality, we assume to have k stations only, and we neglect the remaining ones. Each station contains m sensors. Each sensor measures the carrier up to t times.

The considered dataset collects the t measurements carried out by the m sensors in the k stations over n batches or carriers, assuming a batch to account for a couple of wafers flowing through the whole production line.

So we wrap all the available data into a structured dataset represented by a matrix X with n rows and $y := k \times m \times t$ columns.

As our approach is totally data-driven, without losing generality and for the scope of the model, hereinafter we assume $k = 7$ and $m = 6$. Moreover, we assume $t = 3$, because each sensor measures the carrier three times while it is inside the considered station.

3. Monte Carlo Based Pre-Preprocessing

In this section we illustrate a novel pre-processing approach based on Monte Carlo (MC) simulation and compare it with a commonly used method based on the Interquartile Range (IQR). This last is considered as a reference and the goal is to prove that our approach is a valid alternative to the IQR method. Since both these methods concern only the outlier removal phase, we also briefly describe the preliminary pre-processing steps required to standardize the data and handle missing values or flat signals.

3.1. Preliminary Data Cleaning

Independently on the method, a preliminary data cleaning and preparation stage is required before removing outliers. The following steps are applied:

- signal filtering when the missing values are above 5% of the total number of measurements. Above this threshold, data interpolation can lead to distortions so we preferred to discard the involved signals.
- linear interpolation of signals when the missing values are less that 5% of the total number of measurements.

- flat signals removal when the derivative is zero for at least 50% of the signal length since constant measurements do not provide any meaningful information.
- signal standardization in order to make the scales of the different signals comparable. This operation was achieved by subtracting the mean value and dividing by the standard deviation.

In the next sections we describe the reference IQR method, followed by the discussion of the proposed approach based on MC simulation.

3.2. IQR Method

The Interquartile Range (IQR) method is a simple but effective method used to identify outliers by isolating samples below the 25th percentile or above the 75th percentile [25].

3.3. Monte Carlo Method

In this paper we propose an innovative method for removing outliers based on MC simulation, which has been largely applied in other scenarios like estimation of sum, linear solvers, image recovery, matrix multiplication, low-rank approximation, etc. [26]. In our case, the idea is to generate new data points providing a more robust dataset by applying an estimator to random samples extracted from the original dataset.

By using the median estimator, there is no need to remove outliers from the raw data since this estimator is proved not to be affected by outliers [27].

Moreover, the size of the estimator dataset can be chosen arbitrarily, and can even be greater than that of the original one.

In the next sections we discuss the choice of the proper estimator, the number of samples used for MC simulation and the sliding window approach adopted to preserve the temporal locality of the sensor signals. Finally, we present the pseudocode illustrating the general pre-processing approach used to generate the new estimator dataset as input to the PCA model.

3.3.1. Mean Versus Median

The mean and the median are considered to be the most reliable estimators of the central tendency of a frequency distribution. Choosing the appropriate estimator is a challenging issue when using MC simulation since different results can lead to different correlations between signals, and thus different principal components when applying PCA. Let

$$x_i = (x_{p,z,w})_{\substack{p\ =\ 1,\ldots,k \\ z\ =\ 1,\ldots,m \\ w\ =\ 1,\ldots,t}} \tag{1}$$

denote the i-th row of the $n \times y$ data matrix X accounting for the measurement of sensor z during phase w in station p relative to batch i. In this way, each column f_j ($j = 1, \ldots, k \times m \times t$) of X describes the temporal evolution of the measurements recorded by a specific sensor in a station during the processing of the batches.

Let $R^{IQR} = [r_{ij}^{IQR}]$ with $i, j \in \{1, \ldots, n\}$, $i \neq j$, $r_{ij}^{IQR} = \frac{\sigma_{f_i f_j}}{\sigma_{f_i}\sigma_{f_j}}$ and $-1 \leq r_{ij}^{IQR} \leq 1$ denote the correlation matrix computed between the columns of the dataset resulting from the IQR pre-processing. Recall that $\sigma_{f_i f_j}$ denotes the covariance between the columns f_i and f_j, whereas σ_{f_i} denotes the variance of the i-th column.

Let $R^{MC,median} = [r_{ij}^{MC,median}]$ and $R^{MC,mean} = [r_{ij}^{MC,mean}]$ $(i, j \in \{1, \ldots, n\}, i \neq j)$, formulated as above, denote the correlation matrix computed between the columns of the dataset resulting from the median-based and the mean-based MC simulation pre-processing methods, respectively.

Let $\Delta := [\delta_{ij}] = R^{IQR} - R^{MC}$ account for the deviation between the two matrices, letting R^{MC} denote alternatively the correlation matrix relative to the median-based or the mean-based MC pre-processing method.

In order to evaluate which estimator suits our purpose best, we run the following statistical hypothesis test:

$$\begin{cases} H_0 : \delta_{ij} < \alpha \quad \forall i, j \\ H_1 : \delta_{ij} \geq \alpha \quad \forall i, j, \end{cases} \tag{2}$$

considering the difference between the correlation matrix computed after the application of the IQR method and the correlation matrix of the new dataset resulting from the previous section (that is, the MC dataset).

We can state that there exists a significance level α such that $\delta_{i,j}^{MC,median} < \alpha, \forall i, j$, and $\exists (i, j) : \delta_{i,j}^{MC,mean} \geq \alpha$, allowing us to choose H_0 only under the median-based MC method.

In particular, in the considered use case, the difference in the correlation matrices considering the median-based MC method is less than $\alpha = 6 \times 10^{-2}$ in absolute value and this proves to be a consequence of the median insensitivity to outlier observations.

3.3.2. Choosing the Size of the Monte Carlo Sample

Choosing the proper number of samples has a significant effect on MC simulation since it considerably improves estimation reliability. We recall that samples are chosen out of the data matrix X, where x_i, as defined in (1), represents a generic row of X accounting for the measurement of sensor z during phase w in station p relative to batch i.

Up to the authors' knowledge, the literature claims that increasing the sample size reduces the variance and decreases the noise of the simulation results method [28]. Calibrating the sample size depends on many factors such as dataset size, the pursued objective and the complexity of the phenomenon the designer is modeling [29]. Therefore, we have tested different sample sizes before defining a methodology aimed at finding a suitable number of samples for each round in MC simulation.

By comparison with the highly dispersed original dataset, by increasing the number of samples we obtain a proportional decrease in variance. The desired sample size will allow to remove only the outliers and at the same time preserve the rest of the information contained in the original dataset.

By excessively increasing the number of samples, the risk is that a significant part of the information is lost, thus affecting the accuracy of the PCA model.

In order to select the proper sample size for MC-based outlier removal, we evaluate the impact this parameter has on the PCA model.

To demonstrate that MC pre-processing is a valid alternative to the IQR-based pre-processing method, we compared the PCA models resulting from both approaches for different sample sizes, ranging from 1 to 100. In particular, we measured the proportion of the variance of the MC-PCA components that is explained by the IQR-PCA components in terms of R^2. In this way, high values of R^2 correspond to similar PCA models, thus confirming the equivalent performance of the two pre-processing methods.

From Figure 2, it is evident that by considering three samples we obtain the highest value of R^2 (around 97.5%), thus demonstrating that, by choosing the proper sample size, the MC pre-processing method achieves very similar results to those obtained by the IQR-based pre-processing method.

Figure 2 presents the results of the previous steps where it is experimentally proven that PCA with three-sample size has the best results.

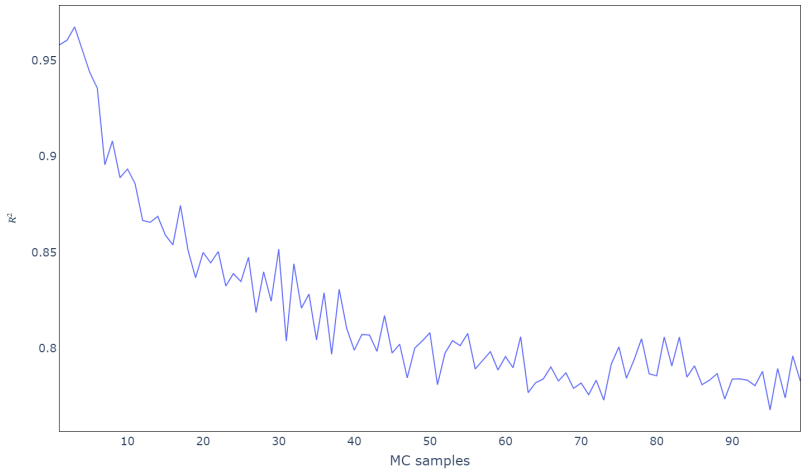


Figure 2. Testing R-squared for different sample sizes.

3.3.3. Preserving Trend Properties through a Suitable Choice of the Monte Carlo Sample

Since PCA is based on the linear correlation among variables, any trends intrinsic to the signals themselves will not be considered. For this reason, a random sampling among all the batches for the purpose of median computation may result in the loss of the temporal dependencies characterizing signals.

Therefore, we refined the procedure for the MC sample selection accordingly. In particular, for each batch in the original dataset we considered a time window centered around the batch itself. Samples considered for the median computation were therefore extracted inside such window, thus preserving the temporal locality among subsequent batches.

3.3.4. Pseudocode for the Pre-Processing Method Based on MC Simulation

The pseudocode reported in Algorithm 1 illustrates the steps required to generate a new estimator dataset by using a pre-processing procedure based on MC simulation, as proposed in Sections 3.3.1–3.3.3.

Algorithm 1 Pre-processing algorithm based on MC simulation

Input X: The original $n \times y$ data matrix
Output $\hat{X}$: The new estimator $\hat{n} \times y$ data matrix
Parameter $\hat{n}$: The size of the new estimator dataset
Parameter b: The number of samples considered for MC simulation
$i \leftarrow 0$
while $i < \hat{n}$ **do**
$idx \leftarrow$ generateRandomInteger$[b, n - b - 1]$
for j in range$[0, y - 1]$ **do**
$window \leftarrow X[idx - b : idx + b, j]$
$\hat{X}[i, j] = \leftarrow$ median$(window)$
end
$i \leftarrow i + 1$
end

4. Principal Component Analysis for Anomaly Detection

Principal Component Analysis (PCA) is a well known method commonly used to reduce the dimensionality of a dataset, by transforming the original set of variables into a smaller one that still contains most of the information in terms of variance. In particular, it is a linear dimensionality reduction method based on Singular Value Decomposition (SVD) that projects the data on a lower dimensional space.

Being $\hat{n}$ the number of samples and let y the number of variables, the $\hat{n} \times y$ data matrix $\hat{X}$ is centered (by removing the mean of every feature) and SVD is applied on

its covariance matrix, thus leading to a subset of orthonormal dimensions, namely the Principal Components (PCs) [30]. Since SVD computes PCs incrementally, their number depends on the pre-defined stopping criterion in searching for the next PC. A common strategy is to define the number of PCs as a function of the minimum variance information to be preserved with respect to the original dataset in order to compress the data sufficiently without loosing too much information.

In this paper we use PCA to perform anomaly detection. For this purpose, it is necessary to isolate a subset of data points associated with a normal behavior of the equipment. This subset is used as input to the PCA algorithm to compute a set of PCs considering as stopping criterion a high variance preservation (at least 90%). Having defined the $y \times z$ projection matrix Π composed by the z PCs, it is now possible to project each data point $\hat{x}_i$ on a lower dimensional space as:

$$c_i = \hat{x}_i \Pi \tag{3}$$

where c_i is the z-dimensional compressed version of $\hat{x}_i$. Then, we transform c_i back to its original space by multiplying it by the the inverse of the matrix Π (being Π orthonormal, the inverse coincides with its transpose), thus obtaining the reconstructed version of the input data:

$$\hat{x}'_i = c_i \Pi^T \tag{4}$$

Finally, we compute the reconstruction error of the sample $\hat{x}_i$ as:

$$e_i = |\hat{x}'_i - \hat{x}_i| \tag{5}$$

where the vector e_i contains the residual of every input feature. Since the model is trained on normal behavior data, the reconstruction error should be low for samples belonging to the same distribution. However, during an anomalous scenario, the error is expected to be high since the associated samples will deviate from such distribution. By considering these vectors as KPIs for the stations in the production lines, it is not only possible to detect anomalies when high errors occur, but also go back to the sensors mostly involved by inspecting the residuals of each single input feature.

Remark 1. *Thanks to the property of outlier replacement, to the median-based approach as introduced in Section 3.3.1, to the optimal choice of the sample size as described in Section 3.3.2 and to the preservation of any temporal dependencies characterizing the input signals as stated in Section 3.3.3, the proposed MC-based pre-processing approach turns out to be a robust alternative to IQR pre-processing. In fact, as it can be seen from the experimental results reported in Section 5, using median-based MC simulation in place of the IQR method for the pre-processing stage yields very similar results, although the number of PCs obtained when applying PCA after MC simulation is slightly higher than the number of PCs obtained when applying PCA after the IQR method.*

Remark 2. *The proposed pre-processing approach based on MC simulation is more adapt to the scenario of energy plants whose data require extensive cleaning. In this respect, if the input data are not cleaned enough, the IQR method, by isolating samples below the 25th percentile or above the 75th percentile, may end up removing a significant part of the original dataset, thus potentially compromising the quality of the subsequent data analytics task. Instead, MC simulation overcomes this obstacle by enabling the data scientist to tune the dimension of the dataset resulting from pre-processing according to the technical specifications of the considered task.*

5. Experimental Results of Anomaly Detection

In the experimental phase, we compared the results of the proposed anomaly detection approach considering both the IQR and MC pre-processing methods. In both scenarios, the relevant data were collected from the MES of the 3SUN Factory and a set of normal behaviour samples was defined for training the PCA model.

5.1. Training and Test Sets

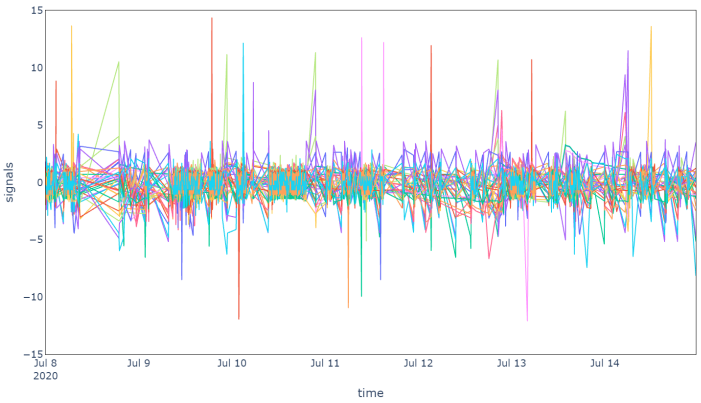
According to the data format of matrix X specified in (1), we isolated a week of normal condition samples as training set, going from 8 July 2020 to 15 July 2020. This period was labelled as a period of standard operation by the operators working in the plant, together with other periods going from 1 November 2020 to 14 November 2020 and from 1 May 2020 to 8 May 2020, respectively, which we considered as test sets. The operators reported a fault in the plant on 4 July 2020, so we isolated 24 days of data before the fault as a further test set to see if the proposed model actually detects the anomaly, possibly in advance.

5.2. Pre-Processing Phase

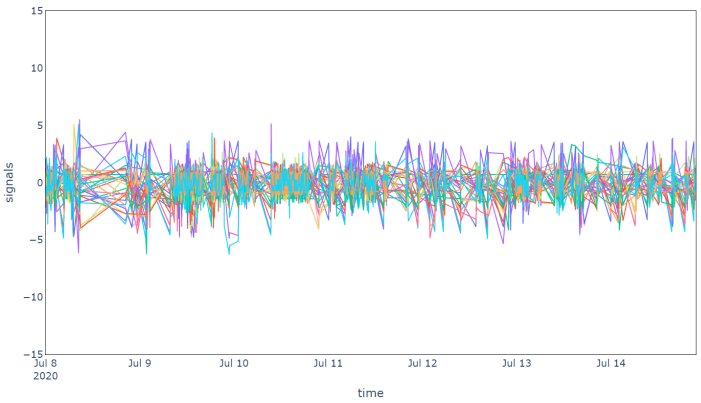
Before the application of the anomaly detection approach based on PCA, we pre-processed the dataset as described in Section 3. In particular, 10 signals were filtered since they were completely flat, 12 signals were discarded since they presented an excessive rate of missing values, and eight signals were linearly interpolated. After this phase, the dataset counted 36 variables on which the two outlier removal methods were applied.

5.2.1. Outlier Removal Results

From the results it is evident that both the IQR and MC methods were able to filter outliers successfully. In Figure 3a, the original sensor signals are plotted in order to highlight the presence of outliers, while in Figure 3b,c, respectively, the pre-processed signals after the IQR and MC outlier removal methods are presented. It is important to notice that the IQR method does not handle the substitution of outliers (e.g., by interpolation) and it is limited to their identification and filtering. The MC method, instead, handles the presence of outliers by replacing all data points with the median over a sliding window, without requiring any additional substitution phase for the filtered values.

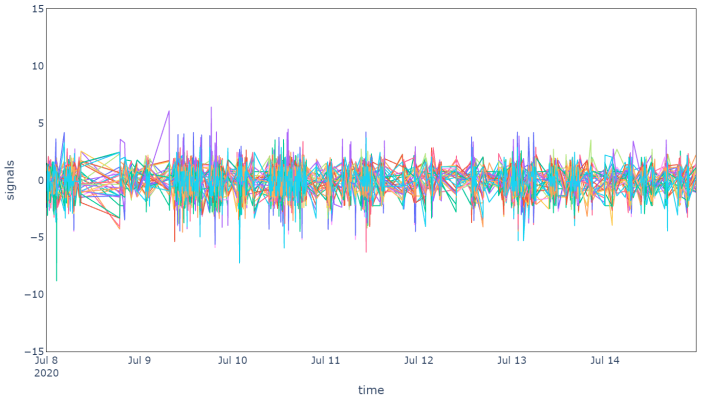


(**a**)



(**b**)

Figure 3. *Cont.*



(**c**)

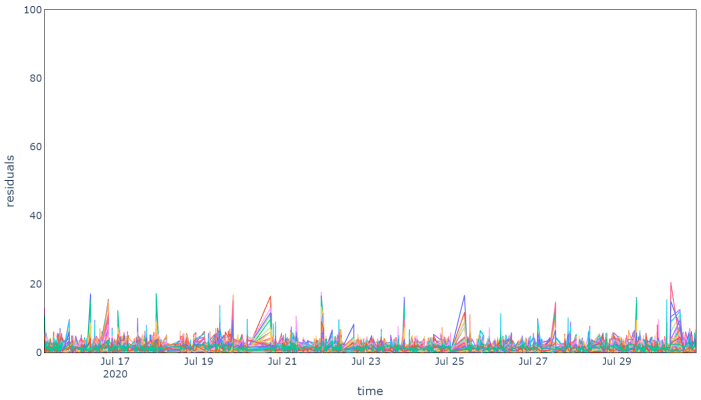
Figure 3. (**a**) shows the sensor signals without the removal of outliers, while (**b**,**c**) represent the signals over time after the IQR and MC methods were applied respectively for the outlier removal phase.

5.3. Anomaly Detection Results

The PCA algorithm was run onto the two scenarios, namely considering an IQR and MC pre-processing phase, by setting as stopping criterion a minimum of 90% of explained variance. In the case of IQR, the PCs computed by the PCA algorithm were 16, while using the MC method led to 19 new dimensions.

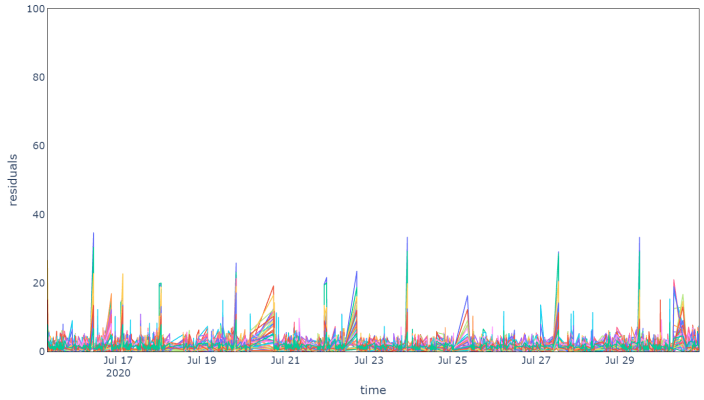
5.3.1. Testing in Normal Operating Conditions

The robustness of the anomaly detection model has been tested on normal behaviour conditions (Figure 4) in a period going from 1 November 2020 to 14 November 2020, namely on the data collected during the week following the training period. Figure 4a plots the reconstruction errors of the model without pre-processing, while Figure 4b,c display, respectively, the residuals considering IQR and MC for pre-processing. In all scenarios the reconstruction errors are never persistently exceeding a threshold of 20 units, which was taken as a reference considering the errors computed on the training data. In fact, the operating conditions are very similar to the normal behaviour period on which the model was trained and demonstrate that there are no substantial differences between the two pre-processing methods.

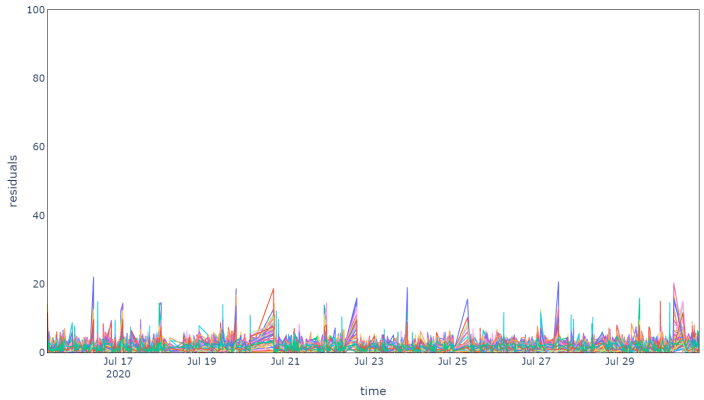


(**a**)

Figure 4. *Cont.*



(**b**)



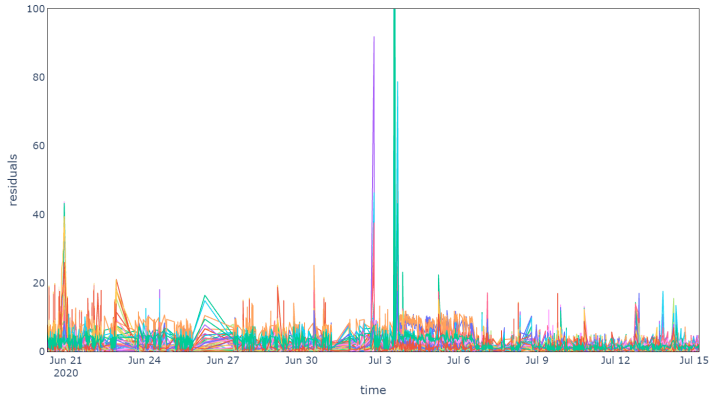
(**c**)

Figure 4. (**a**) shows the KPIs associated to all sensors without the removal of outliers in a normal operating condition period, while (**b**,**c**) represent the KPIs (5) over time after the IQR and MC methods were applied respectively for the outlier removal phase.

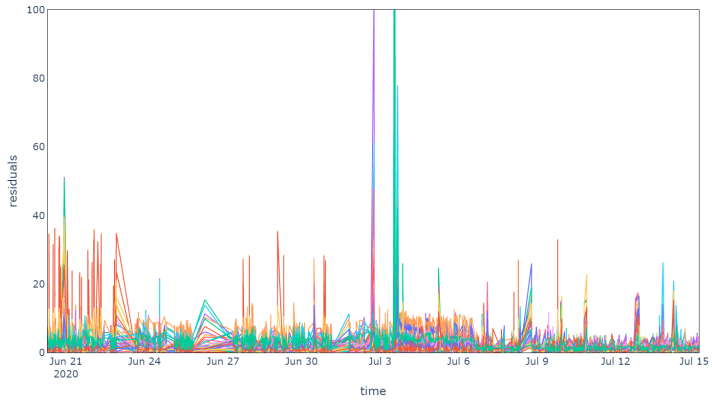
5.3.2. Testing in Anomalous Conditions

As a final step, we evaluated the model in a critical period going from 20 June 2020 to 8 July 2020, during which a technical problem led to equipment failure, as reported by the operators. Figure 5 shows the residuals of the model considering no outlier removal phase (Figure 5a), the IQR (Figure 5b) and the MC (Figure 5c) pre-processing methods. In proximity of the failure event (on 4 July 2020), the anomaly is detected by the residuals drastically exceeding the training reference threshold of 20 units, anticipated by another reconstruction error spike on 3 July 2020. Without outlier removal the residuals never persistently exceed the threshold in the period preceding the fault. When considering the IQR and MC methods, instead, residuals above 20 units are already frequent starting from 20 June 2020, anticipating the fault by more or less two weeks. As for the normal behaviour scenario, also in an anomalous period the two pre-processing methods demonstrated their similarity by achieving comparable results.

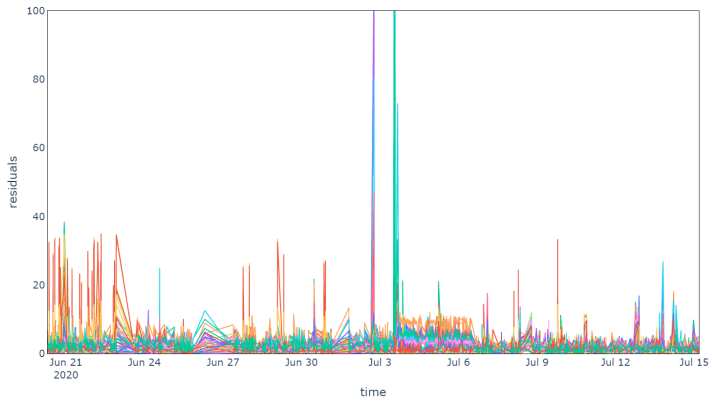
It is important to notice that it is possible to isolate the sensors of the stations that are mostly related to the anomalous conditions by inspecting the residual of each input feature of the model. In this anomalous period, stations 12 and 13 were isolated by looking at the large residuals two weeks before the fault. During the fault itself, instead, stations 19 and 20 were involved according to the model reconstruction errors.



(**a**)



(**b**)



(**c**)

Figure 5. (**a**) shows the KPIs (5) associated to all sensors without the removal of outliers before and after the break, while (**b**,**c**) represent the KPIs (5) over time after the IQR and MC methods were applied respectively for the outlier removal phase.

6. Discussion

The proposed method for data pre-processing based on MC simulation exhibits the following features:

- preserving temporal locality with respect to the training dataset;
- outlier removal;

- outlier replacement, by contrast with traditional methods which are limited to outlier detection only (for example methods based on z-scores [31] or IQR techniques [32]).

As discussed in Section 3.3 and confirmed in [27], the median was chosen as the most accurate estimator in order to obtain a suitable dataset using Monte Carlo simulation to be provided as input to the PCA-based model. In particular, the median-based MC method proved to be more effective against outlier observations with respect to the mean estimator.

Moreover, we selected the optimal sample size for MC simulation by measuring the percentage of variance of the PCA components trained on the MC pre-processed dataset explained by the PCA components trained on the IQR pre-processed dataset in terms of R^2 due to many considerations in the literature which report pre-processing techniques for similar anomaly detection scenarios based on the IQR method [24]. This analysis led to an optimal value of three samples to be considered for the median computation. In particular, we adopted a sliding window sampling approach in order to preserve the temporal locality of subsequent batches.

From the results is Section 5.3 it is evident that the IQR and MC-based pre-processing methods produce similar results, demonstrating their capability to successfully deal with outliers. Nevertheless, they present substantial differences. In fact, a standard method like IQR is limited to isolating outliers and possibly remove them from the dataset. This is a limitation because filtered observations generate missing values which require a substitution algorithm (e.g., mean imputation [33], KNN [34], linear interpolation [35]). The MC method, instead, intrinsically deals with outlier substitution by computing the median of randomly selected points, thus generating a new estimator dataset with an arbitrary number of samples.

The PCA models for anomaly detection demonstrated their capability to successfully anticipate a fault in the equipment as shown in several other works and practical experiments [6–8]. In particular, two PCA models were trained, respectively, on the IQR and MC pre-processed datasets. Both models highlighted an anomalous condition almost two weeks before the equipment failure by producing KPIs (residuals) above a reference threshold which was used to discriminate between healthy and anomalous states of the equipment as done in [36].

Moreover, it is important to notice that, without any pre-processing, the algorithm is unable to detect the anomalies with such an advance and is limited to spotting only the occurrence of the actual fault, which is also detected by the IQR and MC approaches.

Both models were also tested in standard operating conditions in order to prove their robustness to false alarms. In fact, in normal conditions, the residuals of the models never exceed the reference threshold persistently.

Finally, by inspecting the residual of each input feature of the model, the proposed approach allows to isolate the sensors of the stations that are being subject to anomalous conditions.

The authors have selected a reference period in order to calculate the average downtime for the AWB stage of the production line shown in Figure 1, and then to compute an estimate of the AWB downtime reduction resulting from the adoption of our predictive model.

Considering that only 50% of the predicted machine-down events can be totally avoided—in fact, only in some cases it is possible to take advantage of scheduled preventive maintenances to repair the equipment in advance, the authors measured a reduction in AWB downtime by 0.55%. Assuming to extend the implementation of the predictive model to the entire equipment of the 3SUN production line (as shown in Figure 1), the authors expect an overall downtime reduction between 1% and 2%, which corresponds to an increase in the annual photovoltaic panels production in the order of approximately 1–2 megawatts.

7. Conclusions

In this paper, the authors have presented a use case of robust anomaly detection applied to the scenario of a photovoltaic production factory—namely, Enel Green Power's 3SUN solar cell production plant in Catania, Italy—by considering a Monte Carlo based pre-processing technique.

The proposed pre-processing algorithm demonstrated its ability to handle outliers like other standard methods, with the additional advantage of intrinsically dealing with outlier substitution and taking into account the temporal locality of subsequent samples.

After pre-processing, the authors trained an anomaly detection model based on Principal Component Analysis and defined a key performance indicator for each sensor in the production line based on the model errors. In this way, by running the algorithm on unseen data streams, it was possible to isolate anomalous conditions by monitoring the key performance indicators and virtually trigger an alarm when exceeding a reference threshold.

The proposed approach was tested on both standard operating conditions and an anomalous scenario. In particular, it successfully anticipated a fault in the equipment with an advance of almost two weeks, but also demonstrated its robustness to false alarms during normal conditions.

Finally, given the data-driven nature of the approach and its robustness to outliers and irregular sampling frequencies, this approach could be applied to multiple lines in the production plant. In fact, as future work, the authors look forward to testing the proposed method on multiple pieces of equipment in order to further validate its scalability.

Author Contributions: Conceptualization, E.A. and D.A.I.; Methodology, A.C., R.F., E.S.M., L.R.C. and M.V.; Writing—original draft, E.S.M., L.R.C. and N.A.S.; Writing—review & editing, E.S.M. and L.R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by ELIS Innovation Hub within a Joint Research Project with Enel Green Power S.p.A.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AWB	Automatic Wet Bench
CVD	Chemical Vapor Deposition
IQR	Interquartile Range
KPI	key performance indicator
MC	Monte Carlo
MEC	Manufacturing Execution System
PC	Principle Components
PCA	Principle Component Analysis
PeCVD	Plasma Enhanced Chemical Vapor Deposition
PVD	Physical Vapour Deposition
ROBPCA	Robust PCA
SVD	Singular Value Decomposition
WIS	Singular Value Decomposition

References

1. Wang, H.; Ni, G.; Chen, J.; Qu, J. Research on rolling bearing state health monitoring and life prediction based on PCA and Internet of things with multi-sensor. *Measurement* **2020**, *157*, 107657 [CrossRef]
2. Kimera, D.; Nangolo, F.N. Improving ship yard ballast pumps' operations: A PCA approach to predictive maintenance. *Marit. Transp. Res.* **2020**, *1*, 100003. [CrossRef]
3. Chen, X.; Zhang, B.; Wang, T.; Bonni, A.; Zhao, G. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinform.* **2020**, *21*, 1–20. [CrossRef]

4. Kim, S.; Hur, J. A Probabilistic Modeling Based on Monte Carlo Simulation of Wind Powered EV Charging Stations for Steady-States Security Analysis. *Energies* **2020**, *13*, 5260. [CrossRef]
5. Yoo, J.E.; Rho, M. Large-Scale Survey Data Analysis with Penalized Regression: A Monte Carlo Simulation on Missing Categorical Predictors. *Multivar. Behav. Res.* **2021**, 1–29. [CrossRef]
6. De Benedetti, M.; Leonardi, F.; Messina, F.; Santoro, C.; Vasilakos, A. Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing* **2018**, *310*, 59–68. [CrossRef]
7. Bashir, N.; Chen, D.; Irwin, D.; Shenoy, P. Solar-TK: A Data-driven Toolkit for Solar PV Performance Modeling and Forecasting. In Proceedings of the 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Monterey, CA, USA, 4–7 November 2019; pp. 456–466.
8. Bonacina, F.; Corsini, A.; Cardillo, L.; Lucchetta, F. Complex Network Analysis of Photovoltaic Plant Operations and Failure Modes. *Energies* **2019**, *12*, 1995. [CrossRef]
9. Zhou, F.; Park, J.H.; Liu, Y. Differential feature based hierarchical PCA fault detection method for dynamic fault. *Neurocomputing* **2016**, *202*, 27–35. [CrossRef]
10. Chen, Z.; Li, X.; Yang, C.; Peng, T.; Yang, C.; Karimi, H.R.; Gui, W. A data-driven ground fault detection and isolation method for main circuit in railway electrical traction system. *ISA Trans.* **2019**, *87*, 264–271. [CrossRef]
11. Harkat, M.F.; Kouadri, A.; Fezai, R.; Mansouri, M.; Nounou, H.; Nounou, M. Machine learning-based reduced kernel PCA model for nonlinear chemical process monitoring. *J. Control. Autom. Electr. Syst.* **2020**, *31*, 1196–1209. [CrossRef]
12. Bencheikh, F.; Harkat, M.F.; Kouadri, A.; Bensmail, A. New reduced kernel PCA for fault detection and diagnosis in cement rotary kiln. *Chemom. Intell. Lab. Syst.* **2020**, *204*, 104091. [CrossRef]
13. Niemeier, W.; Tengen, D. Stochastic Properties of Confidence Ellipsoids after Least Squares Adjustment, Derived from GUM Analysis and Monte Carlo Simulations. *Mathematics* **2020**, *8*, 1318. [CrossRef]
14. Zelditch, M.L.; Swiderski, D.L.; Sheets, H.D. *Geometric Morphometrics for Biologists*, 2nd ed.; Academic Press: San Diego, CA, USA, 2012; pp. 189–224.
15. Fang, X.; Zeng, W.; Zhou, Y.; Wang, B. On the total least median of squares adjustment for the pattern recognition in point clouds. *Measurement* **2020**, *160*, 107794. [CrossRef]
16. Parashar, S.; Swarnkar, A.; Niazi, K.R.; Gupta, N. Optimal integration of electric vehicles and energy management of grid connected microgrid. In Proceedings of the 2017 IEEE Transportation Electrification Conference (ITEC-India), Pune, India, 13–15 December 2017; pp. 1–5.
17. Fernando, T.M.L.; Marcelo, L.G.E.; David, V.M.H. Substation Distribution Reliability Assessment using Network Reduction and Montecarlo Method, a comparison. In Proceedings of the 2019 FISE-IEEE/CIGRE Conference—Living the Energy Transition (FISE/CIGRE), Medellin, Colombia, 4–6 December 2019; pp. 1–7.
18. Leger, G. Combining adaptive alternate test and multi-site. In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2015; pp. 1389–1394.
19. Kong, X.; Tong, X. Monte-Carlo Tree Search for Graph Coalition Structure Generation. In Proceedings of the 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Chengdu, China, 17–19 October 2020; pp. 1058–1063.
20. Saracco, P.; Batic, M.; Hoff, G.; Pia, M.G. Uncertainty Quantification (UQ) in generic MonteCarlo simulations. In Proceedings of the 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), Anaheim, CA, USA, 27 October–3 November 2012; pp. 651–656.
21. Garcia-Alfonso, H.; Cordova-Esparza, D.M. Comparison of uncertainty analysis of the Montecarlo and Latin Hypercube algorithms in a camera calibration model. In Proceedings of the 2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA), Barranquilla, Colombia, 1–3 November 2018; pp. 1–5.
22. Chen, Y.; Sun, R.; Borken-Kleefeld, J. On-Road NOx and Smoke Emissions of Diesel Light Commercial Vehicles–Combining Remote Sensing Measurements from across Europe. *Environ. Sci. Technol.* **2020**, *54*, 11744–11752. [CrossRef]
23. Heijungs, R. On the number of Monte Carlo runs in comparative probabilistic LCA. *Int. J. Life Cycle Assess.* **2020**, *25*, 394–402. [CrossRef]
24. Nair, P.; Kashyap, I. Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019.
25. Zwillinger, D.; Kokoska, S. *CRC Standard Probability and Statistics Tables and Formulae*; CRC Press: Boca Raton, FL, USA, 2000; p. 18, ISBN 1-58488-059-7.
26. Ji, H.; Li, Y. Monte Carlo methods and their applications in Big Data analysis. In *Mathematical Problems in Data Science*; Springer: Cham, Switzerland, 2015; pp. 125–139.
27. von Brömssen, C.; Röös, E. Why statistical testing and confidence intervals should not be used in comparative life cycle assessments based on Monte Carlo simulations. *Int. J. Life Cycle Assess.* **2020**, *25*, 2101–2105. [CrossRef]
28. Dongxiao, F.; Chuan, P.; Guoxing, Z.; Rui, Z.; Fang, L.; Zhenhua, D.; Hongliang, M. Research on Simulation Method for Reliability Prediction of Pyrotechnical System. In Proceedings of the 2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan), Jinan, China, 23–25 October 2020; pp. 556–559.

29. Meuleman, B.; Billiet, J. A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Surv. Res. Methods* **2009**, *3*, 45–58.
30. Fadhel, S.; Delpha, C.; Diallo, D.; Bahri, I.; Migan, A.; Trabelsi, M.; Mimouni, M.F. PV shading fault detection and classification based on IV curve using principal component analysis: Application to isolated PV system. *Sol. Energy* **2019**, *179*, 1–10. [CrossRef]
31. Aggarwal, V.; Gupta, V.; Singh, P.; Sharma, K.; Sharma, N. Detection of spatial outlier by using improved Z-score test. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 788–790.
32. Vinutha, H.P.; Poornima, B.; Sagar, B.M. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences*; Springer: Singapore, 2018; pp. 511–518.
33. Donders, A.R.; Van Der Heijden, G.J.; Stijnen, T.; Moons, K.G. A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* **2006**, *59*, 1087–1091. [CrossRef]
34. Malarvizhi, M.R.; Thanamani, A.S. K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev.* **2012**, *5*, 5–7.
35. Noor, N.M.; Al Bakri, Abdullah, M.M.; Yahaya, A.S.; Ramli, N.A. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. In *Materials Science Forum*; Trans Tech Publications Ltd.: Zurich, Switzerland, 2015; Volume 803, pp. 278–281.
36. Parzinger, M.; Hanfstaengl, L.; Sigg, F.; Spindler, U.; Wellisch, U.; Wirnsberger, M. Residual Analysis of Predictive Modelling Data for Automated Fault Detection in Building's Heating, Ventilation and Air Conditioning Systems. *Sustainability* **2020**, *12*, 6758. [CrossRef]