Knowledge Discovery from the Statistical Analysis of On-Site Photovoltaic System Data

Knowledge Discovery from the Statistical Analysis of On-Site Photovoltaic System Data

Gianfranco Chicco, Alessandro Ciocia, Andrea Mazza
Dip. Energia "Galileo Ferraris"
Politecnico di Torino
Torino, Italy
{gianfranco.chicco, alessandro.ciocia,andrea.mazza}@polito.it

Radu Porumb
Power Engineering Faculty
Universitatea Politehnica din Bucuresti
Bucharest, Romania
radu.porumb@upb.ro

Filippo Spertino
Dip. Energia "Galileo Ferraris"
Politecnico di Torino
Torino, Italy
filippo.spertino@polito.it

***Abstract*—This paper presents some novel ideas and findings to assist data analysts and operators in discovering specific situations from the analysis of photovoltaic data collected in the field. The daily time series of solar irradiance and active power production are transformed into probability distribution functions (PDFs), then the skewness of the PDF is assessed as a potential indicator of orientation of the PV system in directions different from South. The relevant PDFs corresponding to bright days are identified by applying a clustering procedure. The combined calculation of skewness and correlation between solar irradiance and active power data is also used to discover specific cases in which high shadowing occurs in particular days of the year and at particular times. The analyses are carried out by using data collected on-site in PV systems installed at different locations.**

Keywords—Photovoltaic system, experimental data, time series, clustering, statistics, skewness.

I. INTRODUCTION

The increasing diffusion or renewable energy has made available huge amounts of data that have to be collected and processed to carry out the relevant analyses. The presence of many data coming from the field at different resolutions over time raises the issue of interpreting the correctness of the data, before sending them to further elaborations. This task has to be carefully carried out by an expert of the domain, able to recognise the specific aspects of the dataset under study. As it is not possible to view all data individually, the expert identifies some consistency checks that can be made automatically by executing appropriate procedures. Some of these checks are common to data analysis techniques and are solved through common tools used in information science, for example:

- detection and removal (or correction) of missing data, addressed with dedicated procedures [1];
- detection and correction of bad data (outliers), which requires data cleansing [2] and further steps to ensure that the data are appropriate after pre-processing [3].

When the dataset is complete and the bad data issue has been addressed, the important point is to assess whether the data available are good enough to extract useful knowledge [4]. For this purpose, the specificity of the problem to be analysed has to be taken into account. For example, not all the PV systems that are located in a given area are South oriented, and in many cases the information on the orientation of the PV panels is missing [5]. In some cases, there are mixed solutions, with multiple orientations at the same measured site, which calls for the establishment of an equivalent PV system. In other PV sites, the measured values correspond to the same orientation of the PV panels. In the latter case, an attempt to identify more information from the measured data on the solar irradiance and active power could be worthwhile.

Probabilistic models are used to identify how multiple the correlation of the solar irradiance at multiple sites can affect the PV power production in the network. In [6] the focus is on applying a spatial model of solar irradiance at clear sky, and the PV power production is calculated through a probabilistic power flow. Spatial correlations of the PV power are considered in [7] for joint probabilistic forecasting of PV power and temperature.

In this paper, data collected from photovoltaic (PV) systems are considered. Weather-related and electrical data are gathered from dedicated sensors and data loggers with a given resolution in time and for a certain time period. The environmental variables can be correlated with each other and with the power production [8].

The characteristics of the time series of some main quantities, such as solar irradiance and active power, are addressed in more detail to extract appropriate knowledge that can be used for enhancing the PV energy performance assessment.

In particular, this paper introduces some novel aspects in the statistical characterisation of the PV systems:

1. The analysis of the statistical characteristics of the measured power, from which it is possible to extract useful information to complete the data of the PV sites when only partial information about the location and the orientation is available.
2. The analysis of possible discrepancies between the measurements of solar irradiance and active power generation, which help discovering the presence of shadow effects that limit the PV production in certain periods of the year and at certain hours of the day.

The next sections of this paper are organised as follows. Section II introduces the ideas of considering the daily time series as a probability distribution for the purpose of determining its statistical parameters, with particular reference to the skewness. Section III summarises the data analysis procedure to pass from the initial time series of the relevant PV system data to the calculations needed to determine the skewness of the probability distributions in bright day conditions, including the possible execution of a clustering algorithm to identify the bright days from the measured data. Section IV contains the results of applications to real

measurements on PV systems. Section V presents an application of statistical calculations to identify shadowing conditions in a PV system by comparing solar irradiance and active power measurements. The last section contains the conclusions.

II. Analysis of the Statistical Characteristics of the Time Series

A. From Time Series to Probability Distribution Function

The main point introduced here consists in considering the daily time series of solar irradiance and active power, and to transform the corresponding data into probability distributions. The rationale is to analyse the statistical properties of the resulting probability distributions for extracting additional knowledge on the orientation of the PV system. This process is applicable, as the values of the daily time series considered at day *d* are null at the beginning and at the end of the day, so that what happens in a day does not depend on previous data.

The transformation of a daily time series into a probability distribution is not a trivial task. The process of constructing the probability distribution function (PDF) is shaped to be similar to the application of multiple extractions in the Monte Carlo method. The result to be obtained is a histogram in which the horizontal axis is partitioned into a specified number of classes, and the vertical axis is composed of a certain number of occurrences. In practice:

- let us denote as *y* the variable considered for the analysis (e.g., solar irradiance or active power). Hence, the value of the variable at the time step *t* of the day *d* is indicated as $y^{(d,t)}$;
- on the *horizontal* axis, the *classes* (or bins) are defined with the width Δt equal to the time step of the time series;
- on the *vertical* axis, a *reference amplitude* $\Delta y^{(d)}$ is defined by dividing the maximum value $y_{\max}^{(d)}$ of the time series for the day of analysis by a given integer number n_{y}, such that:

$$\Delta y^{(d)} = \frac{y_{\max}^{(d)}}{n_{\mathrm{y}}} \tag{1}$$

For each class (i.e., time step t) of the daily time series, the (rounded) integer number $n^{(d,t)}$ is calculated and is interpreted as the number of occurrences of the same time step in the time series

$$n^{(d,t)} = \mathrm{round}\left\{\frac{y^{(d,t)}}{\Delta y^{(d)}}\right\} \tag{2}$$

where $\mathrm{round}\{.\}$ denotes the the integer operator (rounded value with zero decimals). The entries to be used for the PDF are the time steps (each one contributing with $n^{(d,t)}$ occurrences), divided by the sum of the data. All the occurrences are considered together in the definition of the final PDF, following the traditional rules: for each class, the PDF value $\vartheta^{(d,t)}$ is defined by dividing the number of occurrences by the total number of occurrences $n_{\mathrm{y,tot}}$ and by the class width Δt:

$$\vartheta^{(d,t)} = \frac{n^{(d,t)}}{n_{\mathrm{y,tot}} \cdot \Delta t} \tag{3}$$

Fig. 1 shows an example of PDF constructed by taking the solar irradiance values in a bright day for a PV plant with West-orientation. The Moon-Spencer model [9] is used to determine the solar irradiance in a bright day, considering a inputs the location of a PV plant with known geographical coordinates (longitude and latitude) and geometrical orientation (azimuth, tilt angle). The PV plant is fixed (i.e., it has no sun-tracking equipment installed).

The coloured area corresponds to the bars of the histogram with one-minute time step. As needed by definition of the PDF, the total area is unitary.

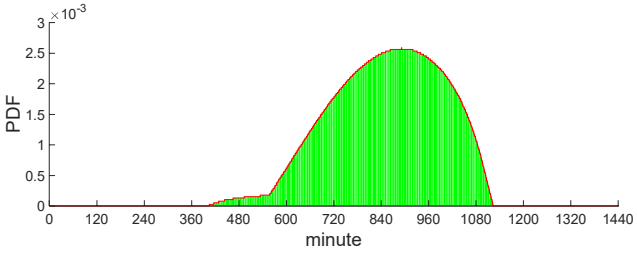


Fig. 1. Example of PDF constructed by using solar model-based irradiance data.

As a numerical example, let us consider the day d = 100 of the year for the PV system that refers to Fig. 1, in which Δt = 1 min and the maximum solar irradiance is $G_{\max}^{(d)}$ = 862.56 W/m^2. By using a partitioning on the vertical axis into n_{y} = 100 levels, the reference amplitude is $\Delta y^{(d)} = G_{\max}^{(d)}/n_{\mathrm{y}}$ = 8.6256 W/m^2. For a given minute (e.g., minute t = 600), the solar irradiance is $G^{(d,t)}$ =224.68 W/m^2, which corresponds to a number of occurrences $n^{(d,t)} = \mathrm{round}\{G^{(d,t)}/\Delta y^{(d)}\}$ = 26. Considering all the minutes during the day, there are $n_{\mathrm{y,tot}}$ =38357 occurrences allocated to the time steps (with zero occurrences during the night). The PDF value at day d = 100 and minute t = 600 is then $\vartheta^{(100,600)} = 6.78{*}10^{-4}$.

B. Probabilistic Characterisation: Skewness

The further action of the probabilistic assessment has been focused on the calculation of the *skewness*, considered as a potentially relevant variable to be associated with the orientation of the PV modules. In particular, with respect to a symmetrical PDF (for which the skewness is null), a left-oriented PDF has positive skewness, while a right-oriented has negative skewness. In terms of the PV systems considered, located in Europe, positive skewness is expected for PV plants with East-oriented PV modules, while negative skewness is expected for PV plants with West-oriented PV modules.

During the year, the solar irradiance has a different maximum value, and the periods from the sunrise to the sunset have different duration. However, with the proposed procedure for determining the PDF, all the PDFs for each day of the year have the same total area (equal to unity, by definition) and are thus fully comparable, without the need of performing further normalisation in amplitude and time as proposed in [10].

The PDFs at the different days of the year can be constructed by using different types of data, among which:

1. The solar irradiance values computed for bright days, taken from an appropriate model, such as the Moon-Spencer model introduced above. The resolution in time can be chosen to fit with the resolution of experimental data, allowing comparative analysis. The daily skewness obtained from these data is denoted as $\chi_{\mathrm{G0}}^{(d)}$.

2. The solar irradiance data measured in a real PV plant, with given resolution in time. In this case, not all the measured data can be considered to be useful for the purpose of determining the PV plant orientation through the calculation of the skewness. Thereby, a *clustering* procedure is applied to the daily measured data to identify the days that can be associated with bright sky conditions, and the skewness is calculated based on measurements gathered during the selected bright days. The daily skewness obtained from the data referring to the selected bright days is denoted as $\chi_{\mathrm{GM}}^{(d)}$.
3. The active power generated by the PV system in bright day conditions, obtained from the solar irradiance values computed for bright days and from other environmental variables assumed at chosen values, using a model of transformation from input variables and further parameters (e.g., efficiency) for obtaining the active power output. The conversion of solar radiation into active power can be done by using an equivalent electrical circuit of a PV generator. The most common equivalent circuit is the Single Diode Model (SDM), that guarantees an optimal compromise between simplicity and high accuracy. Another common model is the Double Diode Model (DDM), which is preferred in case of partial shading [11]. The active power production from a generic PV plant can be also calculated by a model proportional to irradiance and dependent on the temperature of the PV modules. This model is simpler to implement and adequately accurate in case of no shadings on the PV modules [12]. The daily skewness obtained from the active power data referring to the model is denoted as $\chi_{\mathrm{P0}}^{(d)}$.
4. The active power measured in a real PV plant, with given resolution in time. Also in this case, it is necessary to execute a *clustering* algorithm on the daily measured data, to identify the days that can be associated to bright day conditions. The daily skewness obtained from the active power data referring to the selected bright days is denoted as $\chi_{\mathrm{PM}}^{(d)}$.

The overall procedure applied to analyse the data, which includes the statistical characterisation indicated in this section and the execution of the clustering algorithm when needed, is illustrated in the next section.

III. Overall Data Analysis Procedure

A. General Scheme

The data analysis procedure is applied to the daily time series for a one-year duration of the four types of data already indicated, namely:

1. Solar irradiance data for bright days (model-based)
2. Solar irradiance measured on-site (data-driven)
3. Active power generation in bright days (model-based)
4. Active power generation measured on-site (data-driven)

All the data are considered to be complete and with no bad data (i.e., possible data cleansing has already been done).

For every type of data, the first stage is the determination of the daily PDF, using the procedure indicated in Section II.A. The next stage depends on whether the data have been obtained from a model-based approach (with the construction of a model and dedicated simulations) or from a data-driven approach (i.e., from on-site measurements). Fig. 2 shows the computational stages included in the two approaches.

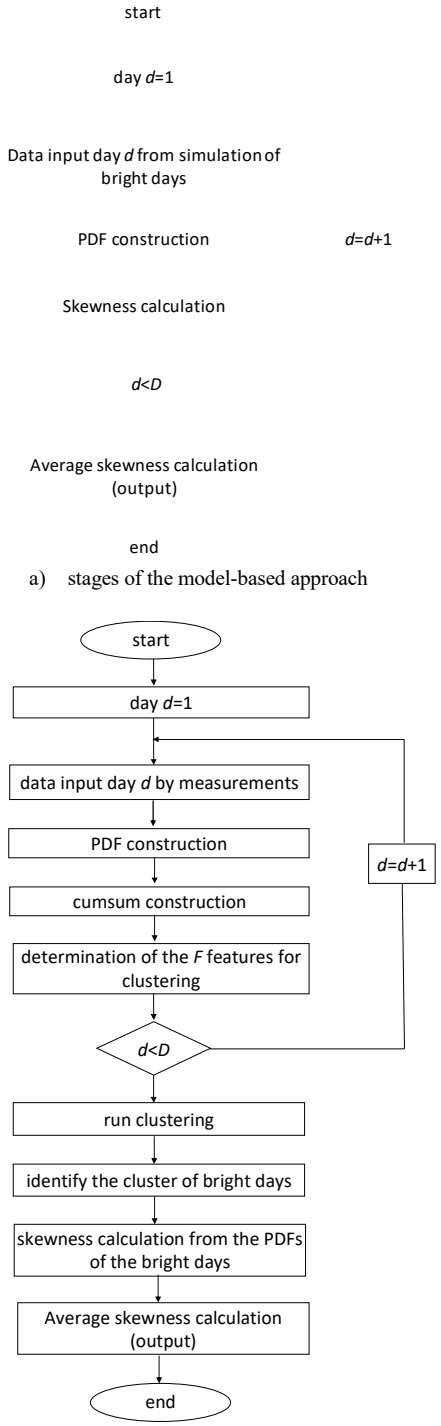


Fig. 2. Flowchart of the proposed calculation procedures.

In particular:

a) From model-based data, the solar irradiance data can be used as they stand, or to construct the model of the PV active power generation. As the data are simulated, it is possible to construct many PDFs referring to different orientations of the PV modules, thus calculating the skewness in each case.

b) From the data-driven approach, the measured values could be gathered in different ambient conditions, e.g., during bright days, but also during cloudy and

intermediate days. The values obtained far from bright conditions are unsuitable to be considered for the analysis of the PDF corresponding to the daily time series. For this purpose, a clustering algorithm is executed for finding out the subset of days that can be considered as bright days (in comparison with the other days).

The identification of the data used for clustering and the choice of the clustering algorithm are discussed below.

B. *Clustering of the Day Types*

The clustering algorithm creates a given user-defined number J of groups (clusters) starting from a matrix of data that in the specific case considered in this paper contains D rows (equal to the number of the days of the year analysed) and F columns (equal to the number of features chosen).

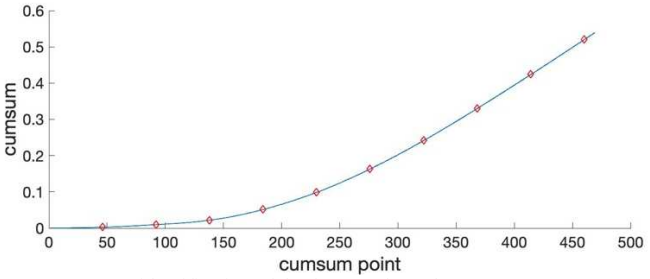
The choice of the features is crucial for making the execution of the clustering procedure effective. Even though high-resolution data can be measured or generated, e.g., at one-minute [13], if all the available data are considered as features, there are $F = 1440$ features, which are too many and too variable [14] to be useful for clustering purposes. In addition, the data can be variable in successive minutes in different days, however, the application of a distance metric such as the Euclidean distance that compares the values at the same minute could not be fully relevant to make a distinction among the types of days. On these bases, the number of features to be chosen has to be reduced.

Moreover, it is important to use the same number of features for all days. However, the number of data available in the time series from the sunrise to the sunset is different during the year. Even considering the number of minutes in the longest daylight period of the year would have the time series exhibiting non-zero data in periods with different duration, and there would be distances between values gathered at the same minute even in bright days. This makes the days non-comparable, unless a suitable normalisation of the horizontal axis is carried out, followed by the definition of refined (e.g., interpolated) time series with the same number of points used as features [10].

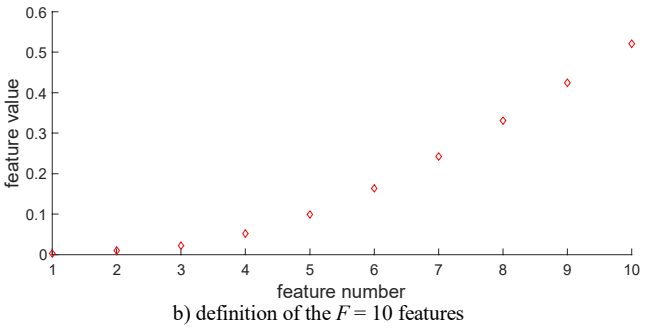
Finally, the maximum solar irradiance (and the corresponding active power) that can be reached in a day depends on the day of the year. Again, comparability among the time series of the various days requires suitable normalisation of the amplitudes, as indicated in [10].

Based on the above issues, in this paper the features are determined with a different rationale. The amplitudes of the daily time series are divided by the maximum daily value. Then, the null amplitudes are removed. The resulting values are divided by the number of time steps with non-null daily values and are ordered in the ascending order. The cumulative sum (cumsum) operator is applied to obtain an ordered set of data in which the relevance of higher values is emphasizes (by the contribution that these values have in the cumsum operator). In this way, there are less than the 1440 features corresponding to all time steps for each day, as there are various time steps with null values.

From the cumsum results referring to all days, the number of points that appear on the horizontal axis is different, and the values are different as well, as no normalisation has been applied. To avoid horizontal axis normalization issues and reduce the number of features at the same time, for each day the portion of the horizontal axis from the minimum to the maximum value is divided into F equal ranges, and the cumsum values found at the upper limits of these ranges are selected as features. In this way, a bright day will have higher values for all the features, while a cloudy day will have lower values. Fig. 3a shows an example of construction of the features for $F = 10$, based on the cumsum data that correspond to active power measurements. The cumsum has 469 points. The locations of the features are at successive positions at $1/F$ of the total number of points (truncated at the lower integer, i.e., at 46 points to each other), and the features are the cumsum values at these locations (marked with diamonds in the figure). Fig. 3b shows the F final features $\mathbf{z}_{\mathrm{PM}}^{(d)} = \{z_{\mathrm{PM},f}^{(d)}\}$, for $f = 1,\ldots, F$.



a) identification of the features from the cumsum



b) definition of the $F = 10$ features

Fig. 3. Determination of the features from the cumsum points.

About the clustering algorithm, in this paper the classical kmeans algorithm is used, because of its general properties of obtaining relatively uniform clusters. Exploring the use of other clustering algorithms is outside the scope of this paper. In general, different executions of the kmeans algorithm lead to different results, unless the seed for random number extractions is fixed before running the algorithm to ensure repeatability of the results.

From the clustering results, the clusters are formed, however, further interpretation of the results is needed to identify the cluster that contains the bright days, as the number of that cluster cannot be established a priori. For the automatic determination of the cluster that contains the bright days, the centroids of the clusters are determined as the average values of the features of the days that belong to each cluster. For example, the centroid $\mathbf{c}^{(j)} = \{c_f^{(j)}\}$ of the cluster $j = 1,\ldots J$, for $f = 1,\ldots, F$, is determined by considering the days that belong to cluster j, denoted as set $\mathbf{D}^{(j)}$ and composed of $n^{(j)}$ entries, so that:

$$\mathbf{c}^{(j)} = \sum_{d\in\mathbf{D}^{(j)}} \frac{\mathbf{z}_{\mathrm{PM}}^{(d)}}{n^{(j)}} \tag{4}$$

Then, a distance measure has been set up, as the sum of the distances between a reference set of points with values $\mathbf{r} = \{f/F\}$, for $f = 1,\ldots, F$, and the F feature values of the centroid $\mathbf{c}^{(j)}$ of the cluster $j = 1,\ldots J$:

$$\delta^{(j)} = \sum_{f=1}^{F} \frac{\left|\frac{f}{F} - c_f^{(j)}\right|}{F} \quad (5)$$

The cluster for which the distance $\delta^{(j)}$ is the lowest is chosen as the cluster that contains the bright days.

The results obtained on an exemplificative case study are reported in the next section.

IV. Application Examples

The concepts indicated above are applied to the daily time series data gathered for one year in a PV plant with the following characteristics. The PV plant is on the rooftop of a dwelling house in northern Italy (N 45.04, E 7.51), and it was installed in 2015. The PV modules (polycrystalline technology) have a rated power of 235 W and a rated efficiency of about 14.5%; the rated power of the entire generator is 2.115 kW. Modules are west oriented with tilt angle of about 37°: they are applied to the tiles of the roof by a metallic structure. Thus, an adequate heat dissipation is obtained by the air passing between the modules and the tiles. There are no near obstacles creating shadows on the roof (e.g., antennas or chimneys); the only shadows are from near houses in the late afternoons.

A. Determination of the Skewness

At first, the time series data of all days have been transformed into the corresponding PDFs by using the procedure indicated in Section II.A. The skewness of the PDFs has then been determined. Fig. 4 shows the values of the skewness obtained from the model-based data of solar irradiance constructed for all the days (blue line). The mean value is -0.465, with variations in the range -0.572 to -0.380. These negative values indicate that the PV plant should be West oriented. The red points in Fig. 4 correspond to the skewness calculated for the PDFs generated at all days. It is apparent that the values are rather different. However, all the days of the year have been considered, with bright, cloud and intermediate days. This result shows that in general it is not possible to reach a solution for the determination of the skewness based on the data of all days. For this purpose, the kmeans clustering algorithm has been used for grouping the day types, searching for the group that represents the bright days.

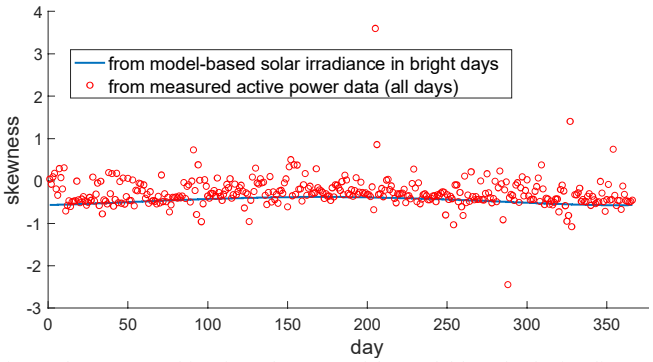


Fig. 4. Skewness resulting from the PDFs of the model-based solar irradiance data and of the measured active power.

The features for clustering have been constructed as indicated in Section III.B. Fig. 5 shows the whole set of days with the associated features (the lines that connect the 10 features for each day are used for easiest representation).

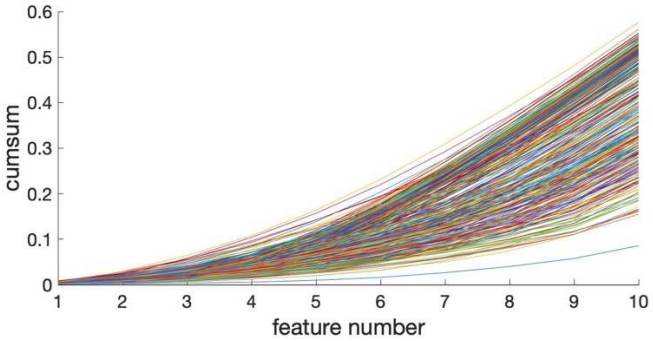


Fig. 5. Clustering input data (features associated to each day).

The kmeans clustering algorithm has been executed with $J = 8$ clusters, considered as a reasonable number of clusters for obtaining meaningful partitioning of the day types in PV applications [10]. Fig. 6 shows the clustering results, in which the J clusters are represented by their features. The initial time series of the corresponding measured active power are shown in Fig. 7. The automatic identification of the cluster with bright days, conducted by using equation (5), indicates cluster #5 as the one where the bright days have been grouped. Looking at Fig. 7, however, it is apparent that at least one uncommon day has been included in that cluster.

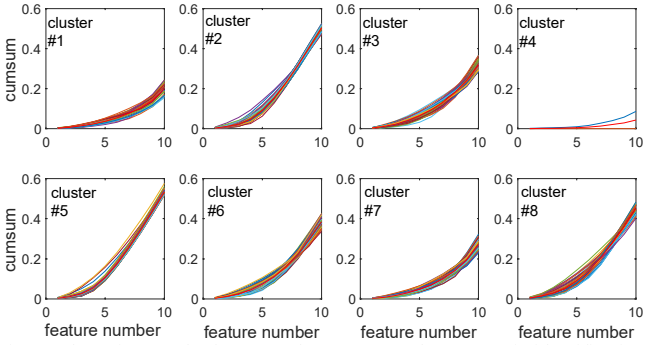


Fig. 6. Clustering results for $J = 8$ clusters. From the automatic identification procedure, cluster #2 contains the bright days.

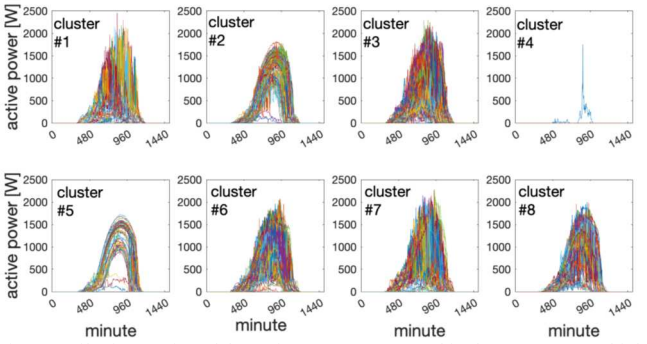


Fig. 7. Daily time series of the active power measured in the PV system which correspond to the clusters created from the features indicated in Fig. 4.

The skewness values of the PDFs that correspond to the days included in cluster #5 are highlighted in Fig. 8 as the blue points. Indeed, the highlighted points correspond to the maximum negative values of the skewness (with the presence of one uncommon value that is associated with the uncommon day indicated above). The mean value of skewness calculated

from the highlighted points is -0.396 and is located inside the range of variation of the skewness determined from the model-based data of solar irradiance. This negative skewness value is consistent with the actual case of West-oriented PV plant.

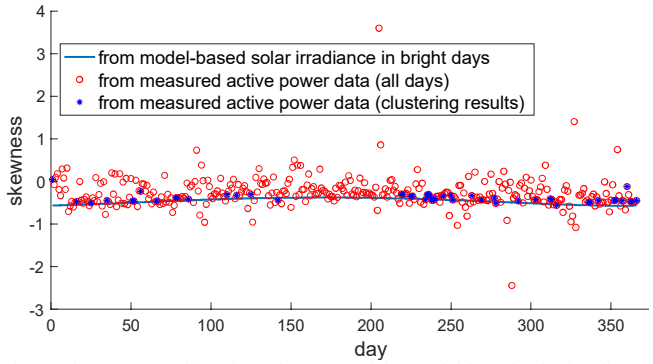


Fig. 8. Skewness resulting from the PDFs of the model-based solar irradiance data and of the measured active power, with indication of the bright days grouped by the clustering algorithm.

From these results, it is confirmed that the determination of the bright days through clustering is a very useful task for obtaining a significant number of days for which the skewness can be calculated. Also, the skewness values are close to the one obtained by using the model-based solar irradiance data in bright conditions.

V. Discovering Shadowing Effects from Measurements

Another example of how pre-processing the measured data may help discovering anomalous operating conditions is described in this section for a PV plant installed in Bucharest, Romania. The solar irradiance and active power have been measured for 56 days in February-March, with time resolution of 5 minutes.

The pre-processing phase includes the calculation of the correlation coefficient between active power and solar irradiance. In the period of analysis, the correlation coefficient resulted in the value 0.533, suspiciously low with respect to normal cases. In fact, in normal cases the correlation between active power and solar irradiance for the same plant should be close to unity [7], because only the non-linearities of the chain from solar irradiance to the active power output affect possible differences in the time series.

In addition, the skewness has been determined by following the procedure indicated in Section III for the data-driven study. Fig. 9 shows the results. The skewness calculated from the solar irradiance measured data is close to zero and exhibits some variations, indicating that the dataset is not formed by bright days only. Moreover, the skewness is close to zero, which could enable guessing that the PV plant could be oriented at or close to the South direction. Conversely, the skewness determined from the active power measurements is rather variable, with both positive and negative values, and in particular is quite different with respect to the corresponding skewness determined from solar irradiance data. This aspect is consistent with the poor correlation found between the measured solar irradiance and active power.

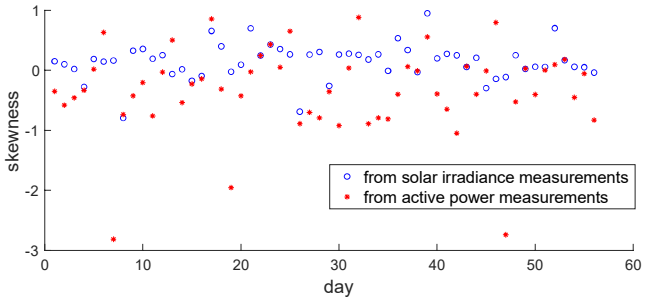


Fig. 9. Skewness determined for the days monitored.

The outcomes arising from the correlation and skewness indicated above indicate the need for investigating possible issues for this PV plant. In the pre-processing phase, the time series of the measured data were not viewed. It is now time to check more data and information on the PV plant.

The PV plant considered for the measurements has a nominal power of 1.2 kW and is South-oriented. Fig. 10 shows the time series for five days in March.

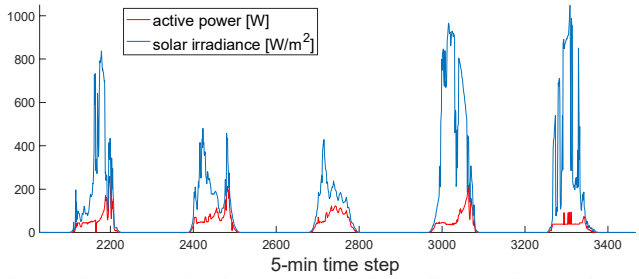


Fig. 10. Time series of active power and solar irradiance with 5-min data for five days.

The discrepancy between the solar irradiance and active power time series is apparent. However, there is no malfunctioning in the measurement systems. The explanation of the measured data has to be found by visiting the installation site to collect more information.

The actual situation is reported in Fig. 11. The skyline in front of the PV panels has high trees, as well as high and far residential buildings (Fig. 11a). The skyline affects the direct

a) skyline

b) mid shadowing

c) heavy shadowing

Fig. 11. Views of the PV plant subject to shadowing in February.

component of the solar irradiance heavily in the winter periods of the year. Fig. 11b and Fig. 11c show some examples of mid and heavy shadowing that occurred mainly in the month of February, confirmed by some photos taken in that period.

The explanation of the discrepancy between measured solar irradiance and active power is that the solar irradiance sensor is located in such a way that is not affected by shadowing as it occurs for the PV panels. Hence, the sensor detects high solar irradiance also when heavy shadowing occurs. This situation, in which the shadows affect the PV panels but not the solar irradiance sensor, is not uncommon in PV plants.

For the PV plant analysed, during the Summer the situation (not shown here) is much better, even though some shadowing could remain in the first period after the sunrise.

VI. Conclusions

This paper has presented some novel findings about the use of statistical analysis to extract knowledge from the data measured on-site in PV systems. The proposed way of pre-processing large amounts of data coming from the PV plant measurements constitutes one of the main steps to add value to raw data. In this way, it is possible to exploit the useful knowledge extracted from the measured data and perform some automatic decisions on the PV plant orientation. Some issues that appear by pre-processing the measured data to provide a meaningful interpretation have been discussed on real cases.

The main results of the analysis are:

- The daily time series of solar irradiance or active power have been transformed into PDFs for making a statistical analysis.
- The skewness of each PDF has been calculated as a potentially useful parameter to describe the PV plant orientation.
- The use of the skewness has been found to be meaningful for bright days, hence a clustering procedure has been set up by defining suitable features to automatically identify a group of bright days, on which the skewness has been calculated.

Further work is in progress to analyse in a systematic way larger PV plant datasets with time series of weather variables and active power production.

References

[1] X. Liu and Z. Zhang, “A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data,” *IEEE Sens. J.*, vol. 21, pp. 10933–10945, 2021.

[2] J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, “Automated load curve data cleansing in power systems,” *IEEE Trans. Smart Grid*, vol. 1, pp. 213–221, 2010.

[3] M. Martinez-Luengo, M. Shafiee, and A. Kolios, “Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation,” *Ocean Eng.*, vol. 173, pp. 867–883, 2019.

[4] G. Chicco, “Data Consistency for Data-Driven Smart Energy Assessment,” *Frontiers in Big Data, section Data Mining and Management*, vol. 4, ref. 683682, 2021.

[5] G. Alba, G. Chicco, A. Ciocia, and F. Spertino, “Statistical Validation and Power Modelling of Hourly Profiles for a Large-Scale Photovoltaic Plant Portfolio,” *Proc. IEEE RTSI 2021*, Naples, Italy, 6-9 September 2021.

[6] J. Widén, M. Shepero, and J. Munkhammar, “Probabilistic Load Flow for Power Grids with High PV Penetrations Using Copula-Based Modeling of Spatially Correlated Solar Irradiance,” *IEEE Journal of Photovoltaics*, vol. 7, no. 6, pp. 1740–1745, 2017.

[7] R. Ramakrishna, A. Scaglione, V. Vittal, E. Dall’Anese, and A. Bernstein, “A Model for Joint Probabilistic Forecast of Solar Photovoltaic Power and Outdoor Temperature,” *IEEE Transactions on Signal Processing*, vol. 67, no. 24, pp. 6368-6383, 2019.

[8] G.G. Kim *et al*., "Prediction Model for PV Performance With Correlation Analysis of Environmental Variables," *IEEE Journal of Photovoltaics*, vol. 9, no. 3, pp. 832–841, 2019.

[9] P. Moon and D.E. Spencer, “Illumination from a non uniform sky,” *Illum. Eng*, vol. 37, pp. 707–726, 1942.

[10] G. Chicco, V. Cocina, and F. Spertino, “Characterization of solar irradiance profiles for photovoltaic system studies through data rescaling in time and amplitude,” Proc. 49th International Universities' Power Engineering Conference (UPEC 2014), Cluj-Napoca, Romania, paper 52, 2-5 September 2014.

[11] N. Barth, R. Jovanovic, S. Ahzi, and M. A. Khaleel, "PV panel single and double diode models: Optimization of the parameters and temperature dependence", *Sol. Energy Mater. Sol. Cells*, vol. 148, pp. 87-98, 2016.

[12] A. Ciocia, A. Amato, P. Di Leo, S. Fichera, G. Malgaroli, F. Spertino, and S. Tzanova, “Self-Consumption and Self-Sufficiency in Photovoltaic Systems: Effect of Grid Limitation and Storage Installation,” *Energies*, vol. 14, no. 6, ref. 1591, 2021.

[13] C.O. Inácio and C.L. Tancredo Borges, "Stochastic Model for Generation of High-Resolution Irradiance Data and Estimation of Power Output of Photovoltaic Plants," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 952–960, 2018.

[14] H.A.H. Al-Hilfi, F. Shahnia, and A. Abu-Siada, "An Improved Technique to Estimate the Total Generated Power by Neighboring Photovoltaic Systems Using Single-Point Irradiance Measurement and Correlational Models," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3905–3917, 2020.