

Statistical Processing of Data Coming from a Photovoltaic Plant for Accurate Energy Planning

Annalisa Di Piazza, Maria Carmela Di Piazza, *Member IEEE*, Gianpaolo Vitale, *Member IEEE*

Consiglio Nazionale delle Ricerche
Institute on Intelligent Systems for the Automation, Palermo – Italy,
(ISSIA – CNR), section of Palermo
Via Dante, 12 90141 PALERMO, ITALY
TEL. +39 091 6113513 FAX +39 091 6113028

mariacarmela.dipiazza@ieee.org, gianpaolo.vitale@ieee.org, annalisadipiazza@hotmail.it

Abstract— This paper presents a statistical approach to manage sampled data coming from a photovoltaic installation. The proposed statistical methods are the *k-means* clustering and the normal density probability distribution. The use of the proposed methods allows to simplify the problem of the PV plant energy assessment respect to the option of obtaining the desired information by managing a large amount of experimental observations. The proposed methods represent useful tools for an appropriate energy planning in distributed generation systems.

Keywords: Photovoltaic energy; Distributed generation; Planning and control of the power system take into account the renewable energy; Models and simulation of the power systems; Software tools.

I. INTRODUCTION

An accurate energy planning in a distributed generation system requires the appropriate knowledge of the renewable energy source capability. When a photovoltaic plant is involved, in particular, the source capability is strictly correlated with the characteristics of the installation site, especially with reference to solar irradiation as it is proportional to the electrical energy deliverable by the PV source. Usually the PV plant energy characterization is based on a long time sampling of main parameters (PV field temperature, solar irradiance, power and energy supplied to the grid etc.). In such a way it is possible to build data bases from which, anyway, it is difficult to extract the information of interest. Therefore the development of analytical tools for the estimation of the quantity of electrical energy generated by a PV plant on a given scale of time is reputed to be very useful. In particular the possibility to find out from the electrical and climatic experimental observations the most significant data to characterise the site of installation from the energy capability point of view is advantageous.

The development of forecasting models for spatial and temporal distributions of climatic variables has been widely treated in technical literature, within the scope of energy assessment.

In such field the synergic use of suitable data processing techniques and estimation methods, either based on statistical or neural approach, represents the more promising way for the set-up of complete and

reliable climatic databases and for the modelling and forecasting of the considered phenomena [1]-[2].

In this paper two statistical tools are proposed in order to obtain an effective estimation of the energy produced by a photovoltaic array in a five-months scale of time, the former being based on the *k-means* clustering methods and the latter on the description of the solar irradiation daily trends through normal probability distributions [3]-[6]. The proposed methods allow to:

1. extract from given scale of time-based experimental measurements the sub-sets of data which can describe the energy capability of the PV plant with good accuracy (*k-means* clustering approach);
2. obtain the energy capability information of the PV plant by describing the solar irradiance trend, on a chosen scale of time, through a continuous function simply defined by two parameters (representation of daily solar irradiance by normal probability distributions).

II. EXPERIMENTAL PLANT

The experimental data have been obtained by a plant installed on a roof footbridge at the University of Palermo – Faculty of Engineering.

The plant has been set up by ENEA (Italian National Agency for New Technologies, Energy and Environment).

The electrical features of the array, under standard test conditions, are the following: open circuit voltage, $V_{oc(stc)}$: 228.2V; short circuit current, $I_{sc(stc)}$: 9.2A; maximum power voltage, $V_{mp(stc)}$: 185.5 V and maximum power current, $I_{mp(stc)}$: 8A.

In Fig. 1 a view of the PV array is reported.

The PV plant is equipped with a data acquisition system that measures the following parameters: panels temperature, solar irradiance, DC voltage and current supplied by the solar array to the inverter, AC voltage, current and power supplied by inverter to the grid [7].

By performing measurements from June to October during all the day and sampling each 10 minutes, a set of more than 12000 couples of experimental values of current and voltage, corresponding to maximum power points (MPPs), for solar irradiance and temperature ranging between 500 and 1100 W/m² and 20 and 50°C, respectively, have been acquired in all.



Fig. 1. View of the PV array

Among these couples only 2295 are situated for solar irradiance and temperature ranging between 500 and 1100 W/m² and 20 and 50°C, respectively.

In [8] a simple clustering of maximum power point data on the basis of solar radiation and temperature has been made. However the problem consisted on the correct choice of irradiance and temperature intervals and, consequently, in identifying the most representative cluster, in terms of energy capability.

III. K-MEANS CLUSTERING METHOD

The k-means clustering is basically a partitioning method. For a given set of observed data the k-means method performs the partition of them into k mutually exclusive clusters. Unlike the hierarchical clustering methods k-means does not create a tree structure to describe the groupings in data, but rather creates a single level of clusters, using the actual observations of objects or individuals in data, and not just their proximities. These features make k-means more suitable for clustering large amounts of data, as in the case under study.

K-means treats each observation in data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. The method of k-means computes cluster centroids, to minimize the sum with respect to a specified measure [3]-[6].

The application of k-means method for the partition of data coming from the studied PV plant has been performed within Matlab[®] environment. In particular the embedded *kmeans* function is used to obtain a vector of indices, indicating to which of the k clusters it has assigned each observation in data and an algorithm, set-up on purpose by the authors, is employed to extract the sets of data assigned to any cluster.

kmeans function uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be further decreased. The result is a set of clusters that are as

compact and well-separated as possible.

To get an idea of how well-separated the resulting clusters are, it is useful to make a silhouette plot using the cluster indices output from *kmeans* function. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. This measure ranges from +1, indicating points that are very distant from neighbouring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. In the described data processing the determination of the correct number of clusters is an issue. A possible way to compare the considered solutions is to look at the average silhouette values for different choices of number of clusters.

In general it is a good idea to experiment with a range of values for k, selecting the value of k corresponding to a partition containing clusters with points having mostly high silhouette values. In this work the described method is used to select the appropriate number of clusters.

IV. NORMAL PROBABILITY DISTRIBUTION

The normal distribution, also referred as the Gaussian distribution, is a widely used continuous probability distribution. The reason for its popularity is due to its attitude to accurately represent many quantitative phenomena in the field of either natural and social sciences. Such distribution is defined by two parameters, i.e. the mean, μ and the standard deviation σ . These parameters represent, respectively, factors of location and scale. In other words, for a given variable x, which can be described by the normal distribution, the mean is its most probable value while the standard deviation is a measure of the spread of x around the mean value. The most common analytical form of the normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1)$$

The normal distribution is also the most widely used family of distributions in statistics.

The sampling distribution of the sample mean approximately follows, for example, the Gaussian curve, even if the observations from which the sample is taken is not normal-distributed. Moreover, the normal distribution maximizes information entropy, i.e. the measure of the uncertainty associated with a random variable, among all distributions with known mean and variance. This feature makes it the natural choice for describing the distribution of data summarized in terms of mean and variance. For the case under study, possible sudden atmospheric variations are filtered if the normal distribution is employed to describe the daily solar irradiance trend. Moreover describing the daily solar irradiance trend by a Gaussian distribution, it is possible to manage a series of sampled data by a continuous function with continuous derivative, simply defined by two parameters. In Fig. 2 the daily trend of the solar irradiance versus time, observed the day October the 3rd, is reported.

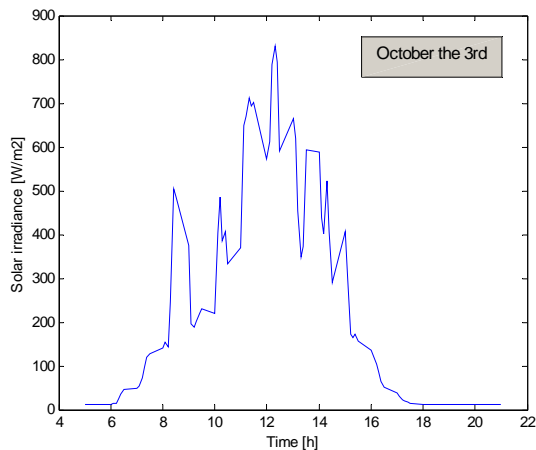


Fig. 2. Solar irradiance versus time, observed the day October the 3rd.

V. ENERGY ASSESSMENT OBTAINED THROUGH K-MEANS CLUSTERING

A k-means-based partition of the overall observed data is performed in order to evaluate the presence of data sub-sets which allow to describe accurately the energy capability of the PV plant. For the scope, all the observed couples of maximum power points voltages and currents, delivered by the plant in the period of observation, have been taken as starting data set.

As a definite knowledge of how many clusters are really in the data, by the energy point of view, is unavailable, the choice for k has been made on the basis of some experiments for k ranging from 3 to 6. In particular the average value h of silhouette plots obtained for each value of k has been evaluated. The best clustering of starting data set has been obtained for k=3. In Figs 3 and 4 the silhouette plots corresponding to k=3 and k=5 are reported respectively. Both plots are obtained minimizing the sum of squared Euclidean distances from centroid for each cluster.

From Fig. 3 it is possible to notice that that cluster 1, in the partition with k=3, contain most of the starting data set. In particular the dimension of cluster 1 is (2x8242), while dimensions of clusters 2 and 3 are respectively (2x2455) and (2x106). So assuming cluster 1 as the most significant data sub-set, the corresponding produced energy during the observations contained in the data of cluster 1, is calculated as:

$$E_{G1} = \sum_{i=1 \div 8242} V_{i1} I_{i1} \Delta t \quad (2)$$

where V_{i1} and I_{i1} are the i^{th} observed data contained in cluster 1 and Δt is the period of each observation (sampling period). This energy amount is then compared with the total energy produced in the five months period of observation, obtaining a deviation of 9.14kWh. Such a deviation corresponds to the 1.2% of total produced energy, equal to 0.75MWh, as $E_{G1}=0.745\text{MWh}$. Therefore it is possible to assess that the proposed clustering method allows to extract, from overall observed data, the most significant group in order to define the PV plant energy capability at the given installation site.

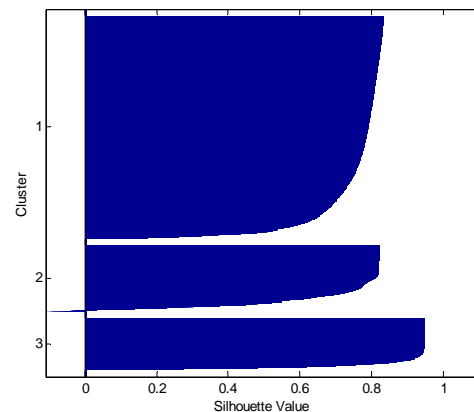


Fig. 3. Silhouette plot for k=3.

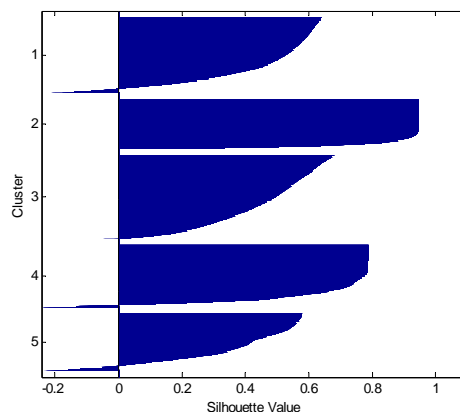


Fig. 4. Silhouette plot for k=5.

This method makes possible an accurate characterization of the PV plant respect to its electrical capability managing only the most significant data sub-set.

VI. ENERGY ASSESSMENT OBTAINED THROUGH REPRESENTATION OF DAILY SOLAR IRRADIANCE BY NORMAL PROBABILITY DISTRIBUTIONS

The whole set of the observed data has been divided, extracting from it the experimental observations related to each daily solar irradiance. Starting from the daily sampled data, normal probability distributions have been obtained, in Matlab[®] environment, by using embedded functions, taken from the Statistic Toolbox. In particular an ad hoc algorithm, using *normfit*, *normplot* and *normpdf* Matlab functions, has been created. Parameter estimates and confidence intervals for the data, supposed to be distributed according to the normal curve, have been obtained on the basis of the *maximum likelihood* method. In such a way the means and the standard deviations for each daily irradiance have been calculated. In Table I the obtained parameters and confidence intervals of the irradiance normal distributions related to two day of observations, i.e. July the 15th and August the 3rd, are reported. Table I contains also the measured supplied energy, E_{md} and the daily supplied energy, E_{cd} , calculated by the mean of irradiance as follows:

$$E_{cd} = \mu_{Gd} \cdot S \cdot T_{obsd} \cdot \eta_d \quad (3)$$

where μ_{Gd} is the mean of the daily irradiance, S is the surface of the PV array, T_{obsd} is the daily period of observation in hours (16 hours, in the studied case) and η_d is the mean daily PV plant efficiency, deduced by the experimental measurements. From table I it is possible to observe that, with the proposed approach, an estimation of the PV plant daily energy capability with a deviation 13% and 9% from the experimental measurement is obtained, respectively, for the two considered days.

In Figs. 5 and 6 the normal probability plots, related to the probability density functions (pdf) of the daily solar irradiance referred in Table I, are shown. Such a plot is useful to graphically assess whether the data in the set of the mean daily irradiance values could come from a normal distribution.

If the data are normal the plot will be linear, otherwise curvature will be introduced in the plot. In the case under study the plot shows a good fitting of the mean daily irradiance values to the normal distribution.

In order to obtain useful information on the energy capability of the PV plant in its installation site for a longer scale of time, a further normal distribution has been deduced by using all the normal daily distributions. Such distribution is obtained using, as stochastic variable, the set of all the mean daily irradiance values and it is related to the whole period of measurements. Therefore it is possible to describe the irradiance trend of all the period of observations through a unique Gaussian curve.

The parameters and the confidence intervals for this new cumulative distribution together with estimated and actual energy capability of the PV plant in the complete period of observation are reported in Table II.

In Fig. 7 the pdf of the cumulative normal distribution, related to the period referred in Table II, is shown.

In Fig. 8 the normal probability plot related to the pdf in Fig. 7 is reported.

With regard to energy assessment, a comparison between energy actually delivered by the plant in the overall period of observation, E_m , with the estimated one, E_c , shows the effectiveness of the proposed method to predict the PV plant energy capability in its installation site. As a matter of fact a deviation of 1.8% between measured and estimated energy is noticed.

The estimated cumulative energy is calculated according to the following relation:

$$E_c = \mu_G \cdot S \cdot T_{obs} \cdot \eta \quad (4)$$

where μ_G is the mean of all the mean daily irradiance value, S is the surface of the PV array, T_{obs} is the whole period of observation in hours (2101 hours, in the studied case) and η is the mean PV plant efficiency, deduced by all the experimental measurements.

The proposed approach is very advantageous also in consideration of the recent diffusion of techniques for the real time laboratory emulation of PV arrays. This approach, in fact, gives the possibility to manage the solar irradiance variations, in different scale of time, by means of a continuous function simply defined by the

two parameters (μ and σ) instead of managing a large amount of sampled data, whose experimental implementation on programmable hardware could be very demanding.

TABLE I
PARAMETERS OF THE DAILY IRRADIANCE NORMAL DISTRIBUTION AND ENERGY EVALUATION

Day	μ_{Gd} [W/m ²]	95% confidence interval for μ_{Gd}	σ_{Gd}	95% confidence interval for σ_{Gd}	E_{cd} [kWh]	E_{md} [kWh]
July15 th	340.83	279.33- 402.32	305.12	267.39- 355.33	4.45	5.12
Aug.3 rd	330.33	274- 386.66	279.48	244.93- 325.48	5.52	6.08

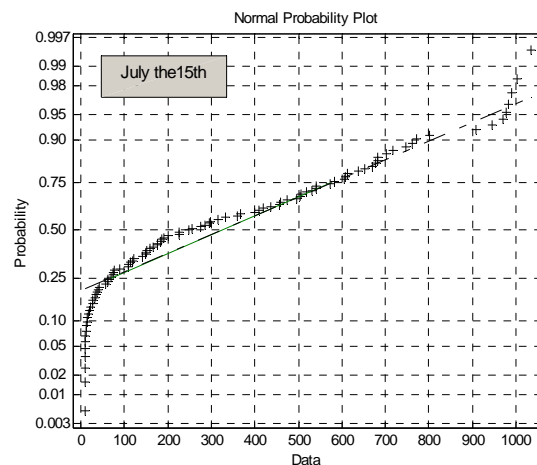


Fig.5 Normal probability plot of the daily normal distribution of solar irradiance referred to July the 15th.

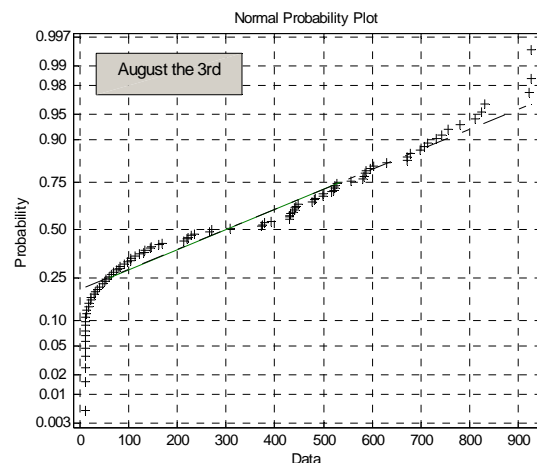


Fig.6 Normal probability plot of the daily normal distribution of solar irradiance referred to August the 3rd.

TABLE II
PARAMETERS OF THE CUMULATIVE IRRADIANCE NORMAL DISTRIBUTION AND ENERGY CAPABILITY EVALUATION

Period	μ_G [W/m ²]	95% confidence interval for μ_G	σ_G	95% confidence interval for σ_G	E_c [MWh]	E_m [MWh]
June- October	358.09	342.60- 373.59	88.57	78.89- 100.98	0.741	0.755

ACKNOWLEDGEMENTS

This work has been funded by the Italian MIUR project n. 211.

REFERENCES

- [1] Jeffrey, Stephen J.; Carter, John O.; Moodie, Keith B.; Beswick, Alan R.; 2001 - Using spatial interpolation to construct a comprehensive archive of Australian climate data, in *Environmental Modelling & Software*, 2001, vol. 16, 309-330.
- [2] Tang, W. Y.; Kassim A.H.M.; Abubakar, S. H.; 1996 - Comparative studies of various missing data treatment methods – Malaysian experience, in *Atmospheric Research*, 1996, vol. 42, 247-262.
- [3] Seber, G.A.F., *Multivariate Observations*, Wiley, New York, 1984.
- [4] Jain, A.K., Dubes, R.C., "Algorithms for Clustering Data", Prentice Hall, 1988, pp. 96-101.
- [5] Davies, D.L., Bouldin, D.W., "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, 1979, pp. 224-227.
- [6] P. Erto "Probabilità e Statistica per le Scienze e l'Ingegneria", McGraw-Hill, 2004.
- [7] M.C Di Piazza, C Serporta.; G. Tine, G. Vitale, "Electromagnetic compatibility characterisation of the DC side in a low power photovoltaic plant", 2004 IEEE International Conference on Industrial Technology " vol. 2, 8-10 Dec. 2004, pp.672 - 677 Vol. 2.
- [8] M.C. Di Piazza, C. Serporta, G. Vitale, "A DC/DC Converter Based Circuit Model for a Solar Photovoltaic Array", 21th European Photovoltaic Solar Energy Conference and Exhibition, 4-8 settembre 2006, Dresda, Germania.

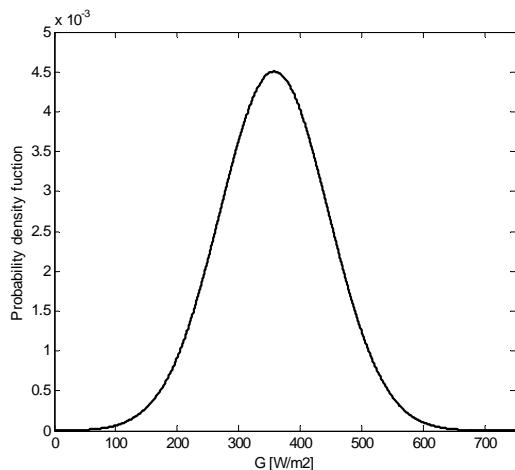


Fig.7 Plot of the cumulative normal distribution of solar irradiance.

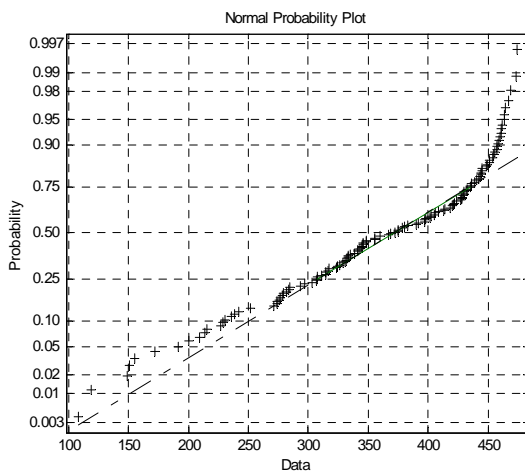


Fig.8 Normal probability plot of the cumulative normal distributions of solar irradiance.

VII. CONCLUSIONS

The use of two statistical approaches is investigated in order to obtain an effective estimation of the energy produced by a photovoltaic array in a five-months scale of time; in particular, the *k-means* clustering methods and the description of the solar irradiance daily trends through normal probability distributions. The application of such statistical approaches to sets of sampled data, coming from an experimental PV plant, allows to simply calculate the plant electrical capability. The *k-means* clustering approach makes possible to extrapolate, from given scale of time-based experimental measurements, the sub-sets of significant data which are sufficient to accurately describe the energy capability of the PV plant.

The method based on the representation of daily solar irradiance by normal probability distributions allows to simply obtain information on the energy capability of the PV plant by describing atmospheric variation trends through continuous functions defined by only two parameters.

The information which can be obtained by using both the two proposed approaches are particularly useful for a suitable energy planning in distributed generation systems.