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, Istituto di Studi sui Sistemi Intelligenti per l'Automazione
(ISSIA – CNR), sezione di 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. Usually the PV plant energy characterization is based on a long time sampling of main parameters. 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 to be reputed very useful.

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 database and for the modelling and forecasting of the considered phenomena [1]-[2].

In this paper two statistical tools are proposed: 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, trough a continuous function simply defined by two parameters (representation of daily solar irradiance by normal probability distributions).

II. 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.

The best clustering of starting data sets has been obtained for k=3. In Fig. 1 the silhouette plot corresponding to k=3 is reported. The plot is obtained by minimizing the sum of squared Euclidean distances from centroid for each cluster.

From Fig. 1 it is possible to notice that cluster 1, contains 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).

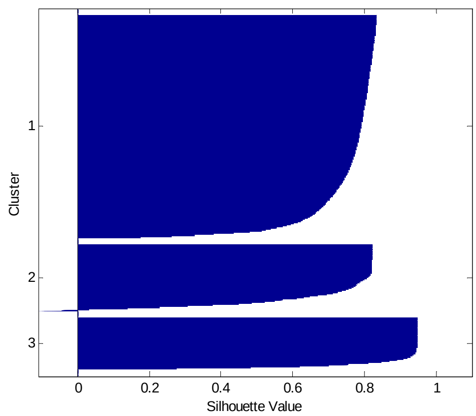

Fig. 1. Silhouette plot for k=3.

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:

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

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 period of observation, obtaining a deviation equal to the 1.2% of total produced energy. 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.

III. Energy Assessment Obtained through Representation of Daily Solar Irradiance by Normal Probability Distributions

Starting from the daily sampled data, normal probability distributions have been obtained, in Matlab® environment. 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 order to obtain useful information on the energy capability of the PV plant a 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 observation 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 I.

Fig. 2 shows the probability density function (pdf) of the cumulative normal distribution, related to the period referred in Table I.

TABLE I
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

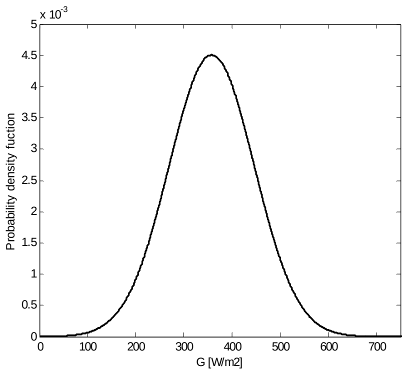

Fig.2 Plot of the cumulative normal distribution of solar irradiance.

With regard to energy assessment, a comparison between energy actually delivered by the plant 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.

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

$$E_c = \mu_G \cdot S \cdot T_{obs} \cdot \eta \qquad (2)$$

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.
As a matter of fact a deviation of 1.8% between measured and estimated energy is noticed.

This approach gives the possibility to manage the solar irradiance variations by means of a continuous function defined by the two parameters (μ and σ) instead of managing a large amount of sampled data.

IV. Conclusions

The use of two statistical approaches is investigated in order to obtain an effective estimation of the energy produced by a photovoltaic; in particular, the *k-means* clustering method and the normal probability distributions have been used. 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 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 trough 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.

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 Exibition, 4-8 settembre 2006, Dresda, Germania.