14th IMEKO TC10 Workshop Technical Diagnostics
New Perspectives in Measurements, Tools and Techniques
for system's reliability, maintainability and safety
Milan, Italy, June 27-28, 2016

Cumulative Statistical Monitoring and Fault Forecasting for PV Plants

S. Vergura

Department of Electrical and Information Engineering, Politecnico di Bari, Italy, silvano.vergura@poliba.it

***Abstract* - This paper proposes a procedure based on statistical tools for diagnosis of PhotoVoltaic (PV) plants. As the data are acquired, statistical analyses are realized. At every new loop other data are added to the previous ones implementing a cumulative statistical analysis. In this way, it is possible to follow the trend of some specific parameters and to understand the real operation of the PV plant, as the environmental conditions change during the year. The proposed approach, based on ANOVA and Kruskal-Wallis tests, is effective in locating abnormal operating conditions. The proposed algorithm has been applied to a real case and results are presented.**

I. INTRODUCTION

Studies on photovoltaic phenomena have drawn the attention to the problem of Photo-Voltaic (PV) system behavior under varying environmental conditions. The raised crucial problem is the strong dependence of the system response from many extrinsic factors, such as insolation intensity, ambient temperature, cell temperature, air velocity, humidity, cloudiness and pollution. All these factors have to be taken into account for the modelling of a PV plant [1]–[3], for the tracking of the Maximum Power Point (MPP), [4], for the monitoring of the energy performance [5]-[6], for the planning of the day after, for the forecasting purposes [7–9]. As reported in [9], PV generation forecasting methods can be broadly classified into three approaches: a) the Numerical Weather Prediction (NWP)-based forecast [10], which uses the first principles for predicting solar irradiance and PV generation; b) the data-driven statistical approach [11], [12], which includes Auto-Regression (AR)-based models and computational intelligence tools such as Artificial Neural Networks (ANNs) [13]; c) the hybrid one, which combines the NWP-based and data-driven models. This paper is focused on the monitoring activity which allows one to check the correct operation of a PV plant and to prevent the failures by means of a timely interpretation of the detected anomalies. Nevertheless, cheap sensors with too low accuracy are often used. It would be suitable that the standardization international organizations (as ISO, IEC and so on) define the minimum values of accuracy for the monitoring systems of PV plants.

This paper is structured as follows: in Section II the proposed algorithm is exposed, Section III describes the PV plant under test and finally Section IV presents the results of the cumulative statistical analysis.

II. ANALYSIS AND DESCRIPTION OF THE PROPOSED PROCEDURE

The random variability of atmospheric phenomena affects the available irradiance intensity for photovoltaic generators. During clear days an analytic expression for solar irradiance can be defined, whereas it is not possible for cloudy days. Extreme-case conditions are usually assumed as reference, for the sake of check, without considering the inherent random nature of some aspects affecting the electrical characteristics of a PV system. The statistical approach for analyzing and monitoring the PV plants allows to take into account the variability of these aspects. In this paper the PV plant will be considered to be composed of k identical sub-arrays, each of them being equipped with a unit of measurement. Each unit will storage measurements of energy produced by the corresponding array, whereas the central monitoring equipment will acquire the total amount of produced energy. The entire set of measures from each sub-array will be considered as a statistical population.

Matlab distribution plots give the possibility to identify an appropriate distribution family for the population, and to draw inferences among samples.

An effective tool for inferential purposes is represented by ANOVA [14] under the following assumptions:

a) distributions of the observations have equal variance;
b) distributions of the observations are normally distributed;
c) the sets of the observations are mutually independent.

The well-known effectiveness of this function in statistical applications and its robustness have become a crucial point for the proposed real time diagnosis.

The ANOVA test is known to be robust with respect to modest violations of the first two assumptions, a) and b). The third hypothesis is always verified in our case, because the measures are taken from independent local unit of measurement. In fact, the modern systems to monitor the PV plants have devoted sensors for each array and their measurements are separately collected in a multi-access data logger. The first step of the procedure consists in verifying if the data populations have equal variances, while the successive step consists in verifying the second ANOVA assumption by means of a normal

probability test [15]. The normal probability plot gives information about the range of values, in terms of percentiles, which fall into the normal distribution.

If the above assumptions are verified it is possible to apply ANOVA test in order to obtain the information related to the p-value.

When the data distribution is not normal, a non-parametric test has to be used. Non-parametric tests make only mild assumptions, but they are less powerful than ANOVA test for normally distributed data.

The Kruskal-Wallis test [16-17] is a non-parametric test based on the assumption that the measurements come from a continuous distribution, which is not necessarily a normal distribution. The test is based on the analysis of variance using the ranks of the data values, instead of the data values themselves.

As the data set dimension increases with time, the fault estimation and location will become more accurate. The proposed algorithm exploits statistical information on the produced energy of the sub-arrays and performs a check on the operation of the PV plant.

The procedure is presented in Fig. 1. Homoschedasticity's test verifies condition a), while normal plot verifies condition b).

As in real cases it is impossible that the data belong exactly to a Gaussian distribution and as ANOVA test can be applied also for modest violation of condition b) (and obviously a)), two indexes have to be calculated in order to quantify the divergence of a real distribution from a Gaussian one.

The first one is the skewness of a distribution, defined as

$$\sigma_k = \frac{E\left(x-\mu\right)^3}{\sigma^3} \qquad (1)$$

where μ is the mean of the data x, σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t. The skewness is a measure of the asymmetry of the data around the mean. For

- $\sigma_k = 0$ the data have a Gaussian distribution;
- $\sigma_k < 0$ the data are spread out more to the left of the mean than to the right;
- $\sigma_k > 0$ data are spread out more to the right.

The second index is the kurtosis, a measure of how outlier-prone a distribution is. Since the kurtosis can have several formulas, here we consider the Pearson's kurtosis less 3, then:

$$k_u = \frac{E\left(x-\mu\right)^4}{\sigma^4} \qquad (2)$$

For:

- $k_u = 0$ the distribution is Gaussian;
- $k_u < 0$ the distribution is less outlier-prone than the Gaussian distribution;
- $k_u > 0$ the distribution is more outlier-prone than the Gaussian distribution.

III. THE PV SYSTEM UNDER TEST

The behavior of a 20 kWp photovoltaic grid connected plant, realized in Bari, Italy, in the year 2003, has been analyzed. It is a grid connected plant that injects the energy exceeding the local consumptions into the distribution network.The 132 panels of the plant, shown in Fig. 2, are partitioned in six equal subsets. The nominal power of a single module is 150 Wp whereas the total power amount for a single subset is 3300 Wp. For each subset a 3000 W inverter has been installed. The system faces the south and is sloped at about 44°. The observation period refers to the year 2009 during which the plant has shown a misoperation and, in particular, an overload of the inverter n.1 and an under-load for the inverter n.5. The unbalanced operation, due to installation errors, has led the overloaded inverter to the fault with the consequent out of service of the corresponding sub-array. The proposed monitoring system has shown good performance in evaluating the incorrect operation of the PV plant; the unbalance event has been sensed and isolated, as reported in the following section.

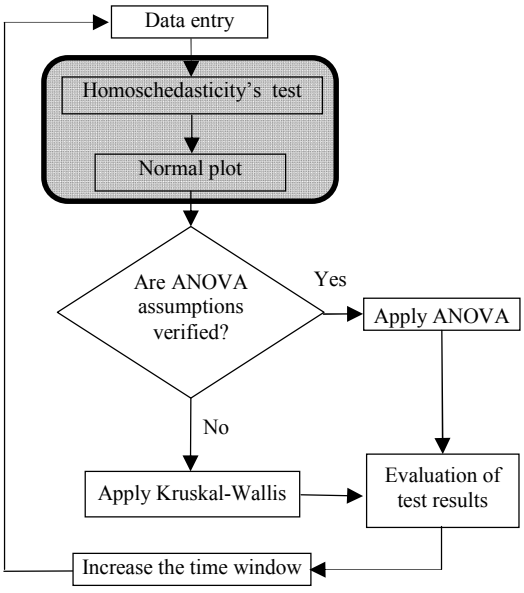

Fig. 1. The proposed algorithm

IV. CUMULATIVE STATISTICAL ANALYSIS

In order to analyze the performance of the PV plant described in Sec. III, statistical tools introduced in Section II have been utilized. Statistical data analysis has been carried out in Matlab R14 environment by using standard routines of the *Statistics toolbox* and by

implementing a new algorithm.

Several analyses are presented in order to evaluate the trends of the energy performances of the PV plant. The increase of the time window described in Fig. 1 allows to understand how some characteristic benchmarks of the PV plants vary during the year as new data are acquired. The incoming analyses represent a cumulative analysis for the statistical monitoring of PV plants. Following results will be presented:

A. 1-month analysis (January 2009);
B. 6-months analysis (January-June 2009);
C. 12-months analysis (January-December 2009).

The following results will be reported for the first analysis: energy produced by each one of the six inverters; normal plot; mean, median, variance and relative spreads of each one of them; ANOVA test and its characteristic parameters; skewness and kurtosis values; *p*-values for ANOVA test. Particularly, values of variance and normal plot are necessary for the first two steps of Fig. 1. For sake of space, the analyses 2-3 will report only normal plots, tables of the values of mean, median, variance and spread, and table of skewness, kurtosis and p-values.

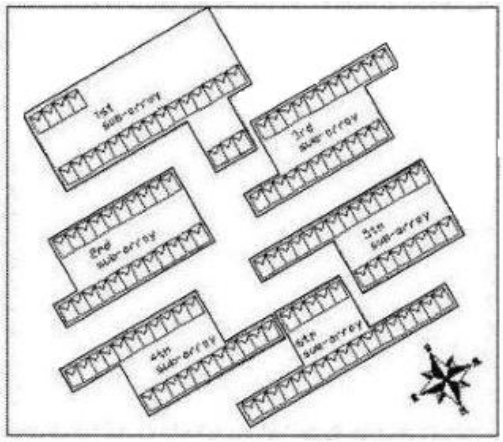

Fig. 2. PV sub-arrays

As real data can never belong to an exactly Gaussian distribution, skewness and kurtosis quantify just the mismatch. Since the thresholds able to unequivocally affirm if data belong or not to a quasi-Gaussian distribution are not fixed in literature, the choice if ANOVA or Kruskal-Wallis has to be applied is uncertain in the cases of little mismatch. In this paper, *p*-value for ANOVA or Kruskal-Wallis test has been evaluated for all the three analyses and a significance level $\alpha = 0.01$ has been pre-fixed. Significance level has to be compared with $p-value$ and $(1-p-value)$.

For the specific case under test, ANOVA conditions were satisfied for the first analysis and not for the last two analyses. Observing Fig. 1, the 3 analyses can be considered as the results of the whole loop for 3 times; in this specific case ANOVA path has been followed for the first analysis and Kruskal-Wallis path for the last two.

A. 1- month analysis (January 2009)

Fig. 3 reports the energy produced by each inverter while Table 1 reports means, medians and variances of the energy produced by each inverter, the global means of them and the spreads in per cent. The spreads of the variances (contained in the range -2.81÷3.46) indicate a modest violation of condition a) of ANOVA test as well as the normal probability plots of Fig. 4 (in which the data belonging to the straight red line are contained in the range 25÷90 percentile) show a modest violation of condition b). Then, applying ANOVA test, box plot (Fig. 5) and its characteristic parameters (Table 2) are obtained. The meaning of columns is the following:

- the first one shows the source of the variability (between columns and within them);
- the second one shows the Sum of Squares (SS) due to each source;
- the third one shows the degrees of freedom (df) associated with each source;
- the fourth one shows the Mean Squares (MS) for each source, which is the ratio SS/df;
- the fifth one shows the *p-value* derived from the cumulative distribution function *F*.

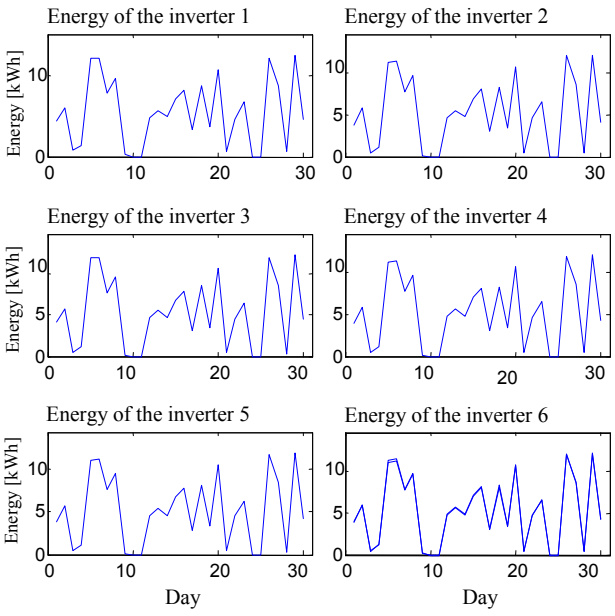

Fig. 3. Energy, in kWh, supplied by each inverters for the 1- month

Table 1. Mean, Median and Variance of the Energy (in kWh) of each inverter and spread with respect to the global values for 1 month

	Inverter number					
	1	2	3	4	5	6
Mean	5.40	5.17	5.25	5.24	5.11	5.19
Global mean	5.23					
Spread %	3.29	-1.04	0.50	0.23	-2.32	-0.66
Median	4.85	4.66	4.66	4.76	4.55	4.75
Global mean	4.71					
Spread %	2.96	-0.94	-0.94	1.23	-3.26	0.96
Variance	17.67	16.88	17.53	17.10	16.60	16.69
Global mean	17.08					
Spread %	3.46	-1.15	2.61	0.13	-2.81	-2.24

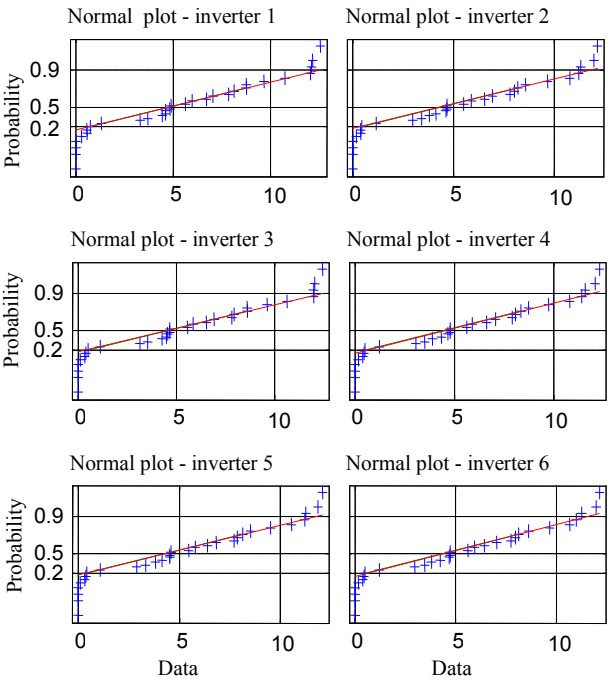

Fig. 4. Normal probability plot for the 6 inverters (1- month)

Table 3 reports the values of skewness and kurtosis in order to quantify the divergence of the six real distributions from the Gaussian ones. Moreover, it reports newly the p-value of ANOVA (0.9999) which implies that the mean values of the six population are almost equal each other (the condition $1-p-value<0.01$ is satisfied for both the tests). Nevertheless, from Table 1, it can be noted that the maximum difference in terms of means spread, equal to 5.61%, and medians spread, equal to 6.22% between inverters 1 and 5 is not small. In most cases the information given by the medians is more effective than that provided by the means. The value of the median mismatch is an alert about the correct operation of the PV plant, even if the p-value affirms the contrary. As all the strings are operating at the same irradiance conditions, the waveforms of Fig. 3 look like almost equal, but the numerical values reveal their differences; moreover, during the days 10, 11, 24 and 25 the PV plant has been stopped for maintenance.

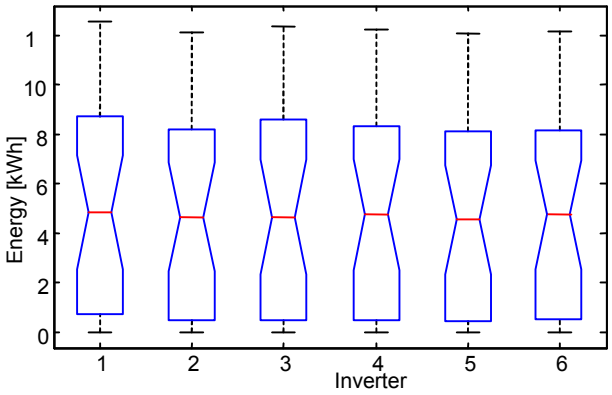

Fig. 5. ANOVA test for the 6 inverters (1-month)

Table 2. ANOVA Values (1- month)

Source	SS	df	MS	**p-value**
Columns	1.48	5	0.2959	**0.9999**
Error	2971.99	174	17.0804	
Total	2973.47	179		

Table 3. Skewness and kurtosis for each inverter (1÷6) p-value of Anova (1-month)

	Inverter number					
	1	2	3	4	5	6
σ_k	0.21	0.20	0.22	0.19	0.21	0.18
k_u	1.85	1.79	1.84	1.79	1.80	1.79
p-value (ANOVA)	0.999					

B. 6-months analysis (January-June 2009)

In this analysis the data of the previous analysis is included. In this case the values of the variance spreads (range [-2.43%÷3.81%] as reported in Table 4) are limited but the data belonging to the straight line of the normal probability plot (Fig. 6) are contained in a small range [25÷75] percentile; then the violation of the condition b) cannot be considered modest and Kruskal-Wallis must be applied. Observing Table 5 it can be noted that the *p*-value (K-W) does not verify the condition $1-p-value<0.01$. Then the hypothesis that the data belong to distributions with the same means must be rejected: at least one population has the mean value different from the others. Table 5 highlights also that the values of skewness have become negative. Coming back to Table 4, it can be noted that the maximum difference

in terms of means spread (equal to 3.33%) regards inverter 6 and inverter 5, whereas the maximum difference in terms of medians spread (equal to 2.88%) regards just the inverters 1 and 5. As pointed already, the median is usually more representative than the mean for the whole population.

Table 4. Mean, Median and Variance of the Energy (in kWh) produced by each inverter and spread with respect to the global values for 6-months

	Inverter number					
	1	2	3	4	5	6
Mean	11.63	11.46	11.36	11.59	11.30	11.68
Global mean	11.50					
Spread %	1.13	-0.40	-1.24	0.78	-1.80	1.53
Median	12.31	12.14	12.04	12.23	11.96	12.27
Global mean	12.16					
Spread %	1.24	-0.18	-0.98	0.59	-1.64	0.97
Variance	40.86	40.81	40.08	41.58	39.98	42.53
Global mean	40.97					
Spread %	-0.27	-0.41	-2.18	1.48	-2.43	3.81

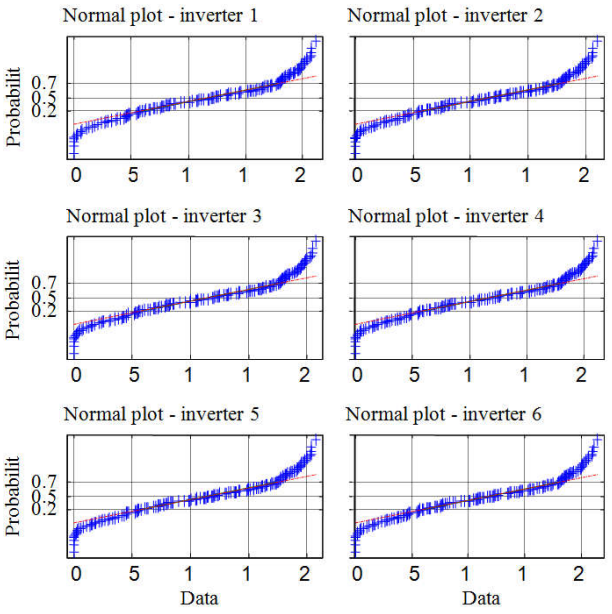

Fig. 6. Normal plot for the 6 inverters (6-months)

Table 5. Skewness and kurtosis for each inverter (1÷6) p-value of Kruskal-Wallis (6-months)

	Inverter number					
	1	2	3	4	5	6
σ_k	-0.26	-0.25	-0.26	-0.25	-0.25	-0.23
k_u	1.75	1.74	1.74	1.74	1.73	1.73
p-value (K-W)	0.955					

C. 12-months analysis (January-December 2009)

This annual analysis contains the whole variability of the environmental conditions of the site in which the PV plant has been set up and then it gives complete information about the overall operation of the PV plant.
In this case the values of the variance spreads (range [-2.42%÷3.84% as reported in Table 6) is limited but the data belonging to the straight red line of the normal probability plot (Fig. 7) are contained in the range [10÷75] percentile; then the violation of the condition b) cannot be considered so modest and Kruskal-Wallis must be applied. Observing Table 7 it can be noted that the *p*-value (K-W) does not verify the condition: at least one population has the mean value different from the others. Table 7 highlights also that the values of skewness are negative and the modules of skewness and kurtosis are similar to those of 6-months analysis.
Table 6 shows that the maximum difference in terms of means spread (equal to 3.09%) regards inverter 6 and inverter 5, whereas the maximum difference in terms of medians spread (equal to 4.82%) regards just the inverters 1 and 5. We still note that the median discriminate better than the mean for the whole population.

From these three analyses it can be noted that an anomaly regarding inverters 1 and 5 is present in the PV plant under examination. This anomaly has been pointed out from the 1-month analysis even if ANOVA has not pointed out it in that analysis. Maybe it is due to the limited amount of the data or to the small violations of condition a) and b). Kruskal-Wallis has been effective in to reveal the anomaly. In fact, an inspection on the plant has allowed to verify that the inverter 1 was upper-loaded, while the inverter 5 was under-loaded. This situation caused several out of orders of the inverter 1 before the anomaly had been detected.

V. CONCLUSIONS

The paper proposes a procedure to statistically analyze the PV plant operation. The procedure is cumulative and some benchmarks are calculated as new data are acquired. Experimental results show only three analyses in order to explain how the procedure is applied during a complete year, but it can be used for real-time monitoring, after specific performance benchmarks have been fixed. In this way it is possible to follow the trend of the benchmarks and to characterize anomalies before they become failures. Alert messages can be sent or a control procedure can be implemented if benchmarks exceed prefixed thresholds. Obviously, the number of applications of cumulative analysis for detecting an anomaly depends on its severity. Real case study has shown the effectiveness of the proposed approach. In fact, the proposed procedure has allowed to reveal

anomalies which have not been detected evaluating the standard indexes. The procedure is suitable to reveal any anomaly or fault which has an effect on the produced energy. Then, overload or under-load conditions as well as specific device faults implying an abnormal production of energy are highlighted. Nevertheless, the procedure does not allow to identify the cause of the fault but only to reveal its effects (on the energy production) and to locate it.

Table 6. Mean, Median and Variance of the Energy (in kWh) produced by each inverter and spread with respect to the global values for 12-months

	Inverter number					
	1	2	3	4	5	6
Mean	22.04	21.78	21.91	21.80	21.30	21.98
Global mean	21.69					
Spread %	1.61	0.41	1.01	0.51	-1.79	1.30
Median	22.32	21.78	22.11	21.45	21.27	21.75
Global mean	21.78					
Spread %	2.48	0.00	1.52	-1.52	-2.34	-0.14
Variance	39.31	39.41	38.60	40.05	38.53	41.00
Global mean	39.48					
Spread %	-0.43	-0.20	-2.23	1.44	-2.42	3.84

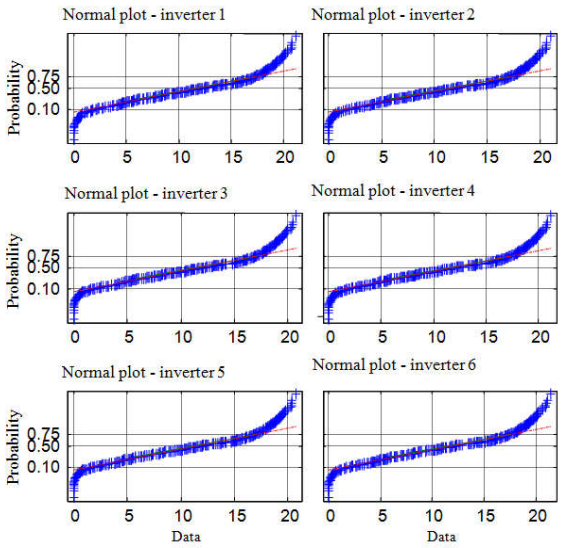

Fig. 7. Normal plot for the 6 inverters (12-months)

VI. REFERENCES

[1] E. Romero-Cadaval, G. Spagnuolo, L. Garcia Franquelo, C.A. RamosPaja,T. Suntio, W.M. Xiao, Grid-Connected Photovoltaic Generation Plants: Components and Operation, Industrial Electronics Magazine, IEEE, 2013, Vol. 7, Issue 3, pp. 6-20.

[2] S. Vergura, "A Complete and Simplified Datasheet-based Model of PV Cells in Variable Environmental Conditions for Circuit Simulation, Energies 9, no. 5: 326, 2016.

Table 7. Skewness and kurtosis for each inverter (1÷6) p-value of Kruskal-Wallis (12-months)

	Inverter number					
	1	2	3	4	5	6
σ_k	-0.36	-0.34	-0.36	-0.34	-0.34	-0.32
k_u	1.83	1.81	1.83	1.81	1.81	1.79
p-value (K-W)	0.873					

[3] S. Vergura, A. Massi Pavan, "On the PV explicit empirical model: operations along the Current-Voltage curve", IEEE-ICCEP 2015 International Conference on Clean Electrical Power, 16–18/06/2015, Taormina, Italy, 2015.

[4] M. Boztepe, F. Guinjoan, G. Velasco-Quesada, S. Silvestre, A. Chouder, E. Karatepe, "Global MPPT Scheme for Photovoltaic String Inverters Based on Restricted Voltage Window Search Algorithm", IEEE Trans. on Industrial Electronics, Vol. 61, Issue 7, 2014, pp. 3302–3312.

[5] L. Cristaldi, M. Faifer, G. Leone, S. Vergura, "Reference String for Statistical Monitoring of Photovoltaic Fields", IEEE-ICCEP 2015 International Conference on Clean Electrical Power, 16–18/06/2015, Taormina, Italy, 2015.

[6] S. Vergura, L. Cristaldi, G. Leone, "Performance Index of Photovoltaic Fields for Diagnostic Purposes", IET-RPG 2016

[7] Y. Hong-Tzer, H. Chao-Ming, H. Yann-Chang, P. Yi-Shiang, "A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output", IEEE Trans on Sustainable Energy, Vol. 5, Issue 3, April 2014, pp. 917–926.

[8] F. Bizzarri, M. Bongiorno, A. Brambilla, G. Gruosso, G.S. Gajani, "Model of Photovoltaic Power Plants for Performance Analysis and Production Forecast", IEEE Trans on Sustainable Energy, Vol. 5, Issue 3, April 2014, pp. 917–926.

[9] C. Yang, A. A. Thatte, L. Xie, "Multitime-Scale Data-Driven SpatioTemporal Forecast of Photovoltaic Generation", IEEE Trans on Sustainable Energy, 2015, pp. 104–112.

[10] J. Zack, "Current status and challenges of solar power production forecasting". In Proceedings of ETWG Solar Workshop, Austin, TX, April 25, 2011.

[11] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting", Solar Energy, vol. 83, no. 10, pp. 1772–1783, 2009.

[12] C. W. Chow et al., "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed", Solar Energy, vol. 85, no. 11, pp. 2881–2893, 2011.

[13] L. Cristaldi, G. Leone, S. Vergura, "Neural Network-Based Diagnostics for PV Plant", IEEE-EEEIC 2016, 07-10/06/2016, Firenze, Italy, 2016.

[14] Hogg, R. V., J. Ledolter, "Engineering Statistics", MacMillan, 1987.

[15] Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, "Graphical Methods for Data Analysis", Wadsworth. 1983,

[16] Gibbons, J. D., "Nonparametric Statistical Inference", 2nd edition, M. Dekker, 1985.

[17]]Hollander, M., and D. A. Wolfe, "Nonparametric Statistical Methods", Wiley, 1973.