European Association for the
Development of Renewable Energies,
Environment and Power Quality (EA4EPQ)

International Conference on Renewable Energies and Power Quality
(ICREPQ'11)
Las Palmas de Gran Canaria (Spain), 13th to 15th April, 2011

Cumulative Statistical Analysis to Monitor the Energy Performance of PV Plants

S. Vergura

Dipartimento di Elettrotecnica ed Elettronica
Politecnico di Bari
Via E. Orabona 4, 70100 Bari (Italy)
Phone/Fax number:+39 080 5963590, e-mail: 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 manner 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 detecting and locating abnormal operating conditions. The proposed algorithm has been applied to a real case and results are presented.

Keywords

Photovoltaic plant, statistics, ANOVA, Kruskal-Wallis.

1. Introduction

In the design of PV plant a crucial problem is the strong dependence of the system response on many extrinsic factors, such as irradiance intensity, ambient temperature, cell temperature, air velocity, humidity, cloudiness and pollution. Successively, when a PV plant has been set up, a monitoring of the system to ensure an optimal performance with respect to the change of environmental conditions is needed.

Standard benchmarks [1], called "final PV system yield", "reference yield" and "Performance Ratio" (PR), are currently used to assess the overall system performance in terms of energy production, solar resource, and system losses. They are defined as follows.

a) Final PV system yield:

$$Y_f = \frac{E}{P_o}[\mathrm{kWh/kW}] \qquad (1)$$

It is the net energy output E divided by the DC power P_0 of the installed PV array. It represents the number of hours that the PV array would need to operate at its rated power to provide the same energy.

b) Reference yield:

$$Y_r = \frac{H}{G}[\mathrm{hours}] \qquad (2)$$

is the total in-plane irradiance H divided by the PV's reference irradiance G. It represents an equivalent number of hours at the reference irradiance. If G equals 1 kW/m^2, then Yr is the number of peak sun-hours. It is a function of the location, orientation of the PV array, and month-to-month and year-to-year weather variability.

c) Performance Ratio (PR):

$$PR = \frac{Y_f}{Y_r} \qquad (3)$$

it is related to the overall effect of losses on the rated output due to: a) inverter inefficiency, wiring, mismatch, and other losses when converting from d.c. to a.c. power; b) PV module temperature; c) incomplete use of irradiance by reflection from the module front surface; d) soiling or snow; e) system down-time; f) component failures.

Unfortunately, they exhibit two drawbacks: a) they supply a rough information about the performance of the overall PV plant; b) they do not allow any assessment of the behavior of the PV plant single parts.

Some authors have considered the use of the statistics for assessing solar PV plant [2], while a monitoring and decision algorithm based on two main theoretical branches of statistical science, namely *descriptive* and *inferential* statistics, has been developed in [3]. The former one is useful to characterize the data population by assigning a proper descriptive model or distribution family to it. The latter one, adopted when the entire set of data is unknown, consists of a data producing process trying to infer the behavior of the entire population from a sub-set of sample data. The idea at the basis of the procedure in [3] is to predict mis-operation events whatever the amount of field measurements is.

This paper proposes an algorithm able to analyze the operation of a PV plant as the data are acquired, implementing a cumulative statistical analysis. This approach allows to monitor also the *trend of some benchmarks* in order to evaluate the operation trend.

For this aim, the algorithm proposed in this paper is based on the whole population of the energy, even if, for a first stage of analysis, it could be applied to sampled data in order to verify if important failures are present.

The paper is structured as follows: Section II introduce the proposed algorithm, Section III describes the PV plant under test and finally Section IV presents the results.

2. Cumulative Statistical Analysis

The random variability of atmospheric phenomena affects the available irradiance intensity for photovoltaic generators. The statistical approach allows to take into account the variability of these aspects.

In this paper the PV plant is 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 will be called *dataset* and is a statistical population.

The whole procedure for the cumulative statistical analysis is reported in Fig. 1.

The first step consists in calculating the standard parameters of the acquired dataset: means, medians and variances of the energy values of the *k* sub-arrays. These values allow to take preliminary information about the correct operating of the PV plant and to highlight strong failures, if present.

The second step consists in evaluating if ANOVA test can be applied or Kruskal-Wallis has to be considered.

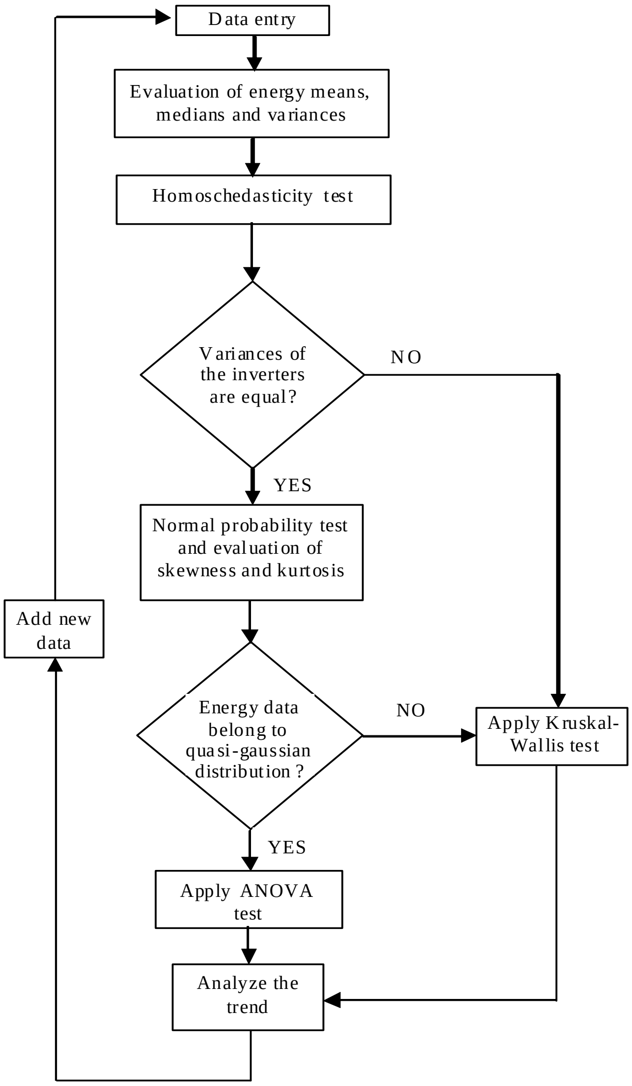
Data entry
Evaluation of energy means, medians and variances
Homoschedasticity test
Variances of the inverters are equal?
NO
YES
Normal probability test and evaluation of skewness and kurtosis
Add new data
Energy data belong to quasi-gaussian distribution ?
NO
Apply Kruskal-Wallis test
YES
Apply ANOVA test
Analyze the trend

FIG. 1 THE PROPOSED ALGORITHM

Both of them are based on a null-hypothesis against the alternative one. The null-hypothesis states that the *k* data populations (one for each sub-array) belong to the same data distribution, i.e. the *k* sub-arrays produce the same amount of energy, if they are constituted by identical PV modules. It implies that the PV plant is working well. Alternative hypothesis is that the *k* population data belong to different data distributions, in spite of they are under the same environmental conditions. Then, if the sub-arrays are constituted by identical PV modules and alternative hypothesis is verified, it implies that operating anomalies or failures are present. In order to decide if null-hypothesis or alternative one is satisfied, it is needed to fix a significance level, with which *p-value* and (1-*p-value*) have to be compared. In fact, if $p\text{-}value<\alpha$ null-hypothesis is rejected, whereas if $(1-p\text{-}value)<\alpha$, alternative one is rejected In this paper a standard value of significance level $\alpha = 0.01$ has been considered.

ANOVA test has well known effectiveness and robustness in statistical applications, if the population satisfy specific constraints. Kruskal-Wallis test [4], instead, is based only on the assumption that the measurements come from a continuous distribution. The test is based on the analysis of variance using the ranks of the data values, instead of the data values themselves (as ANOVA does).

When the constraints of ANOVA are satisfied, ANOVA test gives better results than Kruskal-Wallis. Then, the strategy implemented into the algorithm is:

a) to verify if constraints of ANOVA test are satisfied;
b) if yes, ANOVA test is applied, otherwise Kruskal-Wallis is.

It is needed to keep in mind that ANOVA test can be applied if all the constraints are satisfied; if only one constraint is not satisfied, Kruskal-Wallis has to be considered.

Finally, ANOVA can be used under the following assumptions:

a) all populations have equal variance;
b) all populations are normally distributed;
c) all observations are mutually independent.

The ANOVA test is known to be robust with respect to modest violations of the first two assumptions, a) and b), while the third assumptions is always verified in our case, because the measures are taken from independent local unit of measurement.

In order to verify condition a), homoschedasticity test is applied and then normal probability test verifies the condition b). Normal probability plot gives information about the range of values, in terms of percentiles, which fall into the normal distribution.

In real cases it is impossible that the data belong exactly to a Gaussian distribution; moreover ANOVA test can be applied also for modest violation of condition a) and b), then 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

$$\sigma_k = \frac{E(x-\mu)^3}{\sigma^3} \qquad (4)$$

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.

Fig. 2 reports the case $\sigma_k < 0$, whereas for $\sigma_k > 0$ mode and mean are exchanged respect to the median.

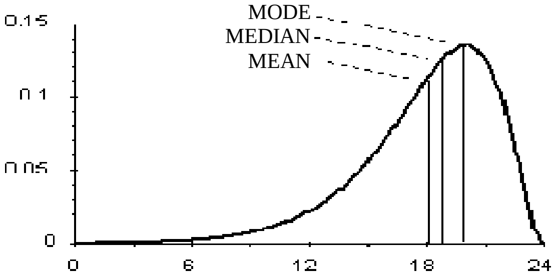

FIG. 2 DATA DISTRIBUTION WHEN $\sigma_k < 0$

The second index is the kurtosis (Fig. 3), a measure of how outlier-prone a distribution is, which is defined as:

$$k_u = \frac{E(x-\mu)^4}{\sigma^4} \qquad (5)$$

For:

- $k_u = 0$ the distribution is Gaussian;
- $k_u < 0$ the distribution is less outlier-prone than the gaussian distribution and is named *platykurtic*;
- $k_u > 0$ the distribution is more outlier-prone than the gaussian distribution and is named *leptokurtic*.

From a mathematical point of view, the *skewness* is a third standardized moment, while the *kurtosis* is a fourth standardized one.

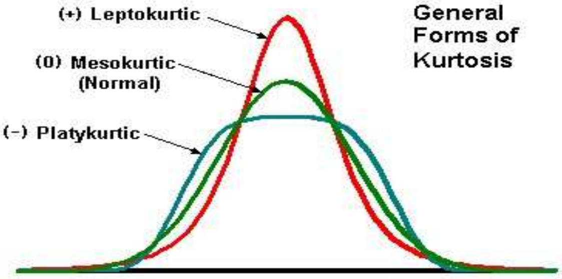

FIG. 3 DATA DISTRIBUTION FOR DIFFERENT KURTOSIS

After the application of the direct flow of the procedure in Fig. 1, following information is collected for a fixed time window:

- energy means, medians and variances (and their spreads in per cent with respect to the global values) for each sub-array;
- normal probability plots of the energy datasets and related skewness and kurtosis for each sub-array;
- p-value for ANOVA or Kruskal-Wallis test and comparison with the prefixed significance level;
- trend of the benchmarks (means and medians spread, skewness).

Then, as new data are acquired, they are added to the previous ones and the direct flow of the procedure is repeated. At each iteration new detailed information is obtained, because of a larger dataset. The incoming analyses represent a cumulative statistical analysis for monitoring the energy performance of PV plants. Cumulative statistical analysis allows to highlight the presence of anomalies in PV plant and to follow the trend of the PV plant operation. Cumulative statistical analysis allows also to locate the anomaly, if present, but it is not able to classify the typology of anomaly or fault neither to define its cause.

3. 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 PV modules are partitioned in six equal sub-arrays. The nominal power of a single module is 150 Wp, while the total power amount for a single sub-array is 3300 Wp. For each sub-array a 3000W inverter has been installed. The system faces the south and is sloped at about 44°. The PV plant is equipped with a datalogger which acquires data from the six inverters. The sample time of the datalogger is fixed to 2 seconds from the manufacturer. After 10 minutes an internal software of the datalogger evaluates the mean of all the measures and only this last value is stored, and so on. The monitoring system measures and stores daily and cumulative values of: a) total power and generated energy on AC side of each inverter; b) voltage Vdc on DC side of each inverter; c) total number of operating hours. The capacity of the monitoring equipment is up to 400 days. The inverter automatically determines the solar generator MPP voltage, which is defined in the internal regulation system as the desired PV voltage.

The observation period refers to the year 2009 during which the plant has shown a mis-operation. The proposed monitoring system has shown good performance in evaluating the incorrect operation of the PV plant; the unbalance event has been sensed, as reported in the following section.

4. Results

In order to analyze the performance of the PV plant described in Sec. III, cumulative statistical analysis introduced in Section II have been applied. The proposed algorithm has been carried out in Matlab R14 environment.

Several analyses will be presented in order to evaluate the trend of the energy performance of the PV plant. The iterative processing of the cumulative dataset allows to understand how some characteristic benchmarks of the PV plants vary during the year.

Following analyses will be carried out:

1) 1-month analysis (January 2009);
2) 3-months analysis (January-March 2009);
3) 6-months analysis (January-June 2009);
4) 12-months analysis (January-December 2009).

The following results will be reported for each analysis: mean, median and variance of the energy values (and relative spreads) for each sub-array; normal probability plots; skewness and kurtosis values; *p*-values for ANOVA or Kruskal-Wallis test; plots of the benchmarks.

A. 1- month analysis (January 2009)

Tab. I 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 (in the range -2.8÷3.5) indicate a modest violation of condition a) of ANOVA test, confirmed by applying homoschedasticity test. Moreover, 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). Tab. II reports the values of skewness and kurtosis in order to quantify the divergence of the six real distributions from the gaussian ones. Then, ANOVA test can be applied. Tab. II reports also the p-value of ANOVA (0.9999); as (1-*p-value*)<*α*, alternative hypothesis can be rejected and it implies that the mean values of the six population are almost equal each other. Nevertheless, from Tab. I it can be noted that the maximum difference in terms of means spread, equal to 5.7%, and medians spread, equal to 6.2% 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.

TAB I
MEAN, MEDIAN, VARIANCE AND SPREAD OF THE ENERGY (IN KWH) OF EACH INVERTER 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	5.10	5.19
Global mean	5.23					
Spread %	3.3%	-1.1%	0.5%	0.3%	-2.4%	-0.7%
Median	4.85	4.66	4.66	4.76	4.55	4.75
Global mean	4.71					
Spread %	2.9%	-0.9%	-0.9%	1.2%	-3.3%	1.0%
Variance	17.67	16.89	17.53	17.10	16.60	16.70
Global mean	17.08					
Spread %	3.5%	-1.1%	2.6%	0.1%	-2.8%	-2.2%

B. 3-months analysis (January-March 2009)

In this analysis the data of the previous analysis (January 2009) are included. As in the previous case, the limited values of the variance spreads (range [-2.3%÷1.7%] as reported in Tab. III) and the normal probability plot of Fig. 5 (the data belonging to the straight line are contained in the range [10÷90] percentile) confirm that ANOVA test can be applied.

Tab. IV reports the values of skewness and kurtosis (similar to the previous ones), while the p-value confirm that the mean values of the six population are almost equal (in fact (1-*p-value*)<*α*). Tab. IV highlights that the maximum difference in terms of means spread (equal to 3.5%) and medians spread (equal to 4.2%) regards just the inverters 1 and 5, even if the per cent mismatch values are decreased.

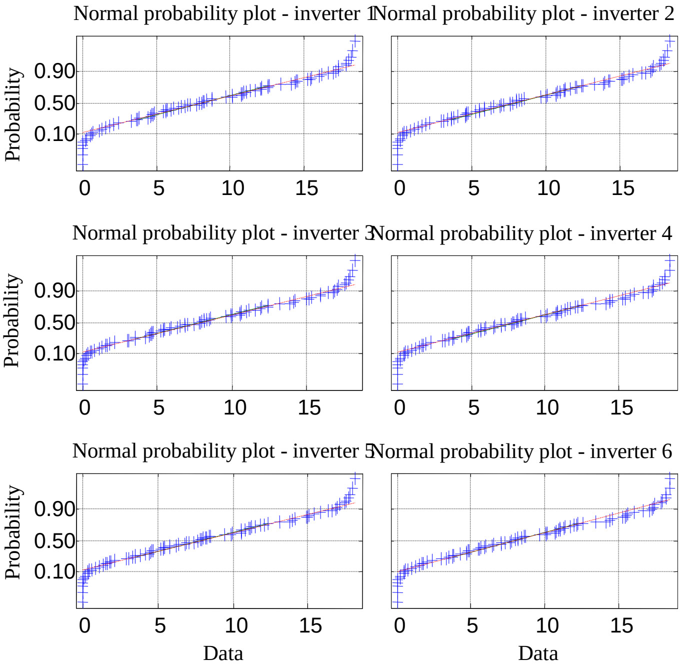

FIG. 4. NORMAL PLOT FOR THE 6 INVERTERS (1- MONTH)

TABLE II. 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)	0.9999					

TAB. III
MEAN, MEDIAN, VARIANCE AND SPREAD OF THE ENERGY (IN KWH) OF EACH INVERTER WITH RESPECT TO THE GLOBAL VALUES FOR 3-MONTHS

	Inverter number					
	1	2	3	4	5	6
Mean	8.39	8.22	8.20	8.31	8.10	8.31
Global mean	8.25					
Spread %	1.6%	-0.4%	-0.7%	0.6%	-1.9%	0.7%
Median	8.10	7.90	7.93	7.96	7.77	7.89
Global mean	7.93					
Spread %	2.2%	-0.3%	0.1%	0.4%	-2.0%	-0.5%
Variance	32.31	32.24	32.10	32.62	31.53	32.80
Global mean	32.268					
Spread %	0.1%	-0.1%	-0.5%	1.1%	-2.3%	1.7%

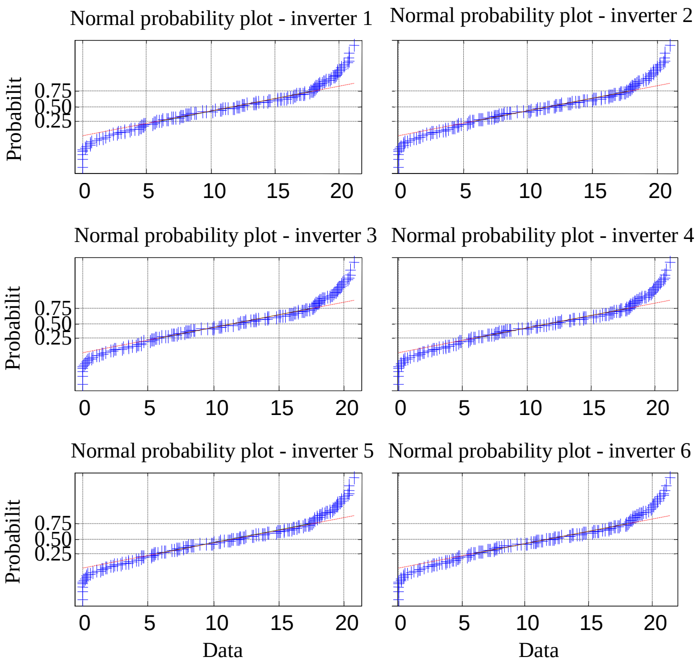

FIG. 5. NORMAL PLOT FOR THE 6 INVERTERS (3-MONTHS)

TABLE IV. SKEWNESS AND KURTOSIS FOR EACH INVERTER (1÷6)
P-VALUE OF ANOVA (3-MONTHS)

	Inverter number					
	1	2	3	4	5	6
σ_k	0.18	0.20	0.19	0.20	0.20	0.21
k_u	1.87	1.88	1.86	1.88	1.87	1.90
p-value (ANOVA)	0.9996					

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

In this analysis the data of the previous two analyses are included. In this case the values of the variance spreads (range [-2.4%÷3.8%] as reported in Tab. V) 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 Tab. VI it can be noted that the *p*-value (K-W) does not verify the condition $1-p-value<0.01$. Then, alternative hypothesis that the data of sub-arrays belong to different distributions (and produce different amount of energy) cannot be rejected.

TAB V
MEAN, MEDIAN, VARIANCE AND SPREAD OF THE ENERGY (IN KWH) OF EACH INVERTER WITH RESPECT TO THE GLOBAL VALUES FOR 6-MONTHS

	Inverter number					
	1	2	3	4	5	6
Mean	11.64	11.46	11.36	11.60	11.30	11.68
Global mean	11.51					
Spread %	1.1%	-0.4%	-1.2%	0.8%	-1.8%	1.5%
Median	12.31	12.14	12.04	12.23	11.96	12.28
Global mean	12.16					
Spread %	1.2%	-0.2%	-1.0%	0.6%	-1.6%	1.0%
Variance	40.86	40.81	40.08	41.58	39.98	42.54
Global mean	40.98					
Spread %	-0.3%	-0.4%	-2.2%	1.5%	-2.4%	3.8%

It implies that at least one population has the mean value different from the others. Tab. VII highlights also that the values of skewness have become negative.

Coming back to Tab. V, it can be noted that the maximum difference in terms of means spread (equal to 3.3%) regards inverter 6 and inverter 5, whereas the maximum difference in terms of medians spread (equal to 2.8%) regards just the inverters 1 and 5. As pointed already, the median values are usually more representative for the whole population than the mean values.

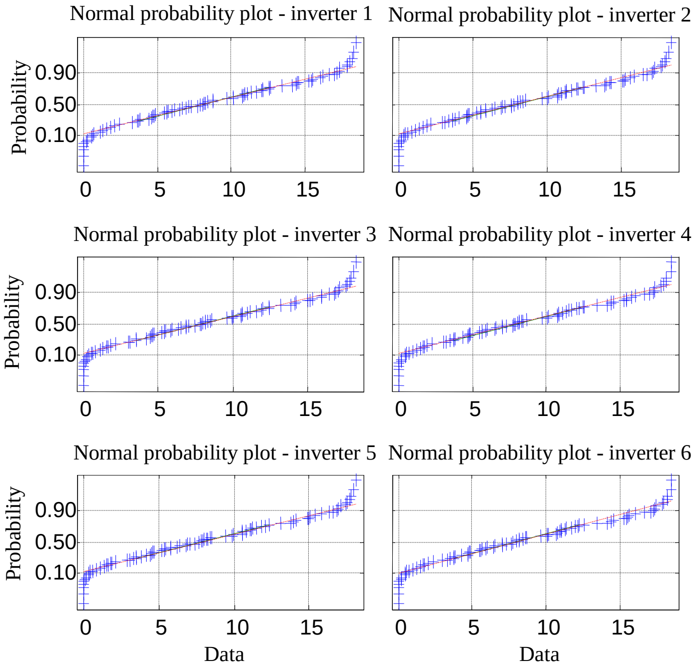

FIG. 6. NORMAL PLOT FOR THE 6 INVERTERS (6-MONTHS)

TABLE VI. 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	-0.26	-0.26	-0.25	-0.25	-0.24
k_u	1.75	1.74	1.75	1.74	1.74	1.74
p-value (K-W)	0.9550					

D. 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 [-2.3%÷3.0% as reported in Tab. VII) 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 Tab. VIII it can be noted that the *p*-value (K-W) does not verify the condition (1-*p-value*)<*α*: at least one population has the mean value different from the others. Tab. VIII highlights also that the values of skewness are negative and the modules of skewness and kurtosis are similar to those of 6-months analysis. Tab. VII shows that the maximum difference in terms of means spread (equal to 3.0%) regards inverter 6 and inverter 5, whereas the maximum difference in terms of medians spread (equal to 3.2%) regards just the inverters 1 and 5. We still note that the median discriminate better than the mean for the whole population.

From these four 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 two cases 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 it had been detected.

TAB VII

MEAN, MEDIAN, VARIANCE AND SPREAD OF THE ENERGY (IN KWH) OF EACH INVERTER WITH RESPECT TO THE GLOBAL VALUES FOR 12-MONTHS

	Inverter number					
	1	2	3	4	5	6
Mean	17.34	17.11	17.22	17.06	16.94	17.45
Global mean	17.19					
Spread %	0.9%	-0.4%	0.2%	-0.7%	-1.5%	1.5%
Median	12.70	12.40	12.40	12.50	12.30	12.40
Global mean	12.45					
Spread %	2.0%	-0.4%	-0.4%	0.4%	-1.2%	-0.4%
Variance	28.11	26.65	27.65	26.78	27.37	26.70
Global mean	27.30					
Spread %	3.0%	-2.3%	1.4%	-1.8%	0.4%	-1.0%

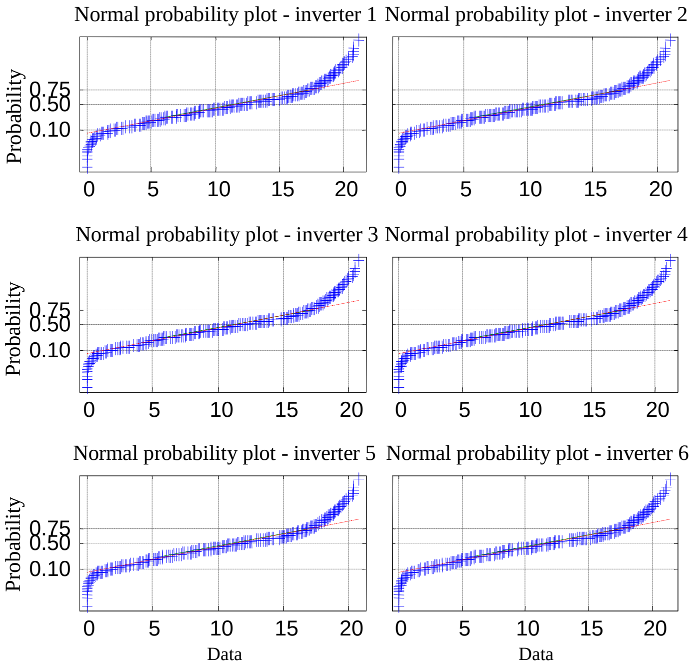

FIG. 7. NORMAL PLOT FOR THE 6 INVERTERS (12-MONTHS)

To evaluate the trend of the operation of the PV plant, Fig. 8 reports the spreads of means (a) and medians (b): sub-array 1 (line blue) results always the maximum value (except for means spread of 3rd and 4th analysis), whereas sub-array 5 (fuchsia line) results always the minimum value. This imply a different operation of sub-arrays n. 1 and n. 5. Fig. 8 (c) reports the trend of skewness: it implies that operation of PV plant has strongly changed after the second analysis (3-months), because skewness has changed its sign. It defines when the anomaly starts.

TABLE VIII. 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	-0.37	-0.34	-0.34	-0.32
k_u	1.83	1.81	1.83	1.81	1.81	1.79
p-value (K-W)	0.8730					

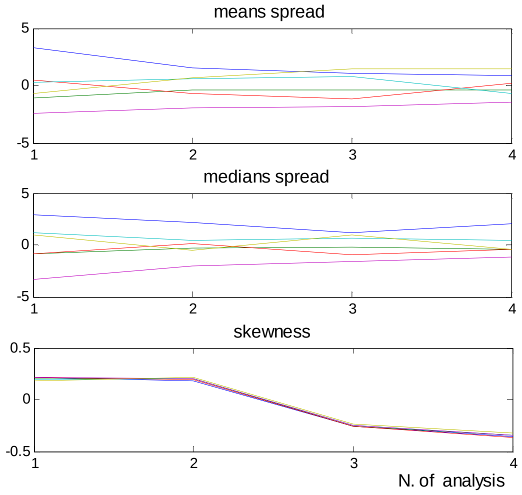

FIG. 8. SPREAD OF MEANS (A) AND MEDIANS (B); SKEWNESS (C)

5. Conclusions

The paper proposes a procedure to statistically analyze the PV plant operation. The procedure is cumulative and benchmarks are calculated and updated as new data are acquired. Experimental results will show only four 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 manner it is possible to follow the trend of the benchmarks and to characterize anomalies before they become failures. Nevertheless, the algorithm give no information about the typology of the anomaly and its cause.

I. REFERENCES

[1] CEI-IEC International Standard 61724- Photovoltaic system performance monitoring- Guidelines for measurement, data exchange and analysis, 1998.

[2] Paul D., Mukherjee D., Bhadra Ch audhuri S.R.,"Assessing solar PV behavior under varying environmental conditions- a statisticalapproach", *in 4th International Conference on Electrical and Computer Engineering ICECE 2006*, 19-21 December 2006, Dhaka, Bangladesh.

[3] S. Vergura, G. Acciani, V. Amoruso, G. Patrono, F. Vacca, "*Descriptive and Inferential Statistics for Supervising and Monitoring the Operation of PV Plants*",IEEE Trans on INDUSTRIAL Electronics (ISSN- 0278-0046), November 2009, pp. 4456-4464.

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