20th IMEKO TC4 International Symposium and
18th International Workshop on ADC Modelling and Testing
Research on Electric and Electronic Measurement for the Economic Upturn
Benevento, Italy, September 15-17, 2014

# Big Data and Efficiency of PV Plants

Silvano Vergura

*Technical University of Bari, Dept. of Electrical and Information engineering*
*st. E. Orabona, 4, 70125 Bari, Italy, silvano.vergura@poliba.it*

***Abstract* – Sometimes many PhotoVoltaic (PV) plants have to be monitored by an unique supervision centre. In these cases the real-time analysis of the whole population of data by means of statistical approach requires long time and it is not always possible. After an introduction on Big Data and statistical approaches, the paper presents a procedure of analysis based on sampled data and on the whole population, only for inefficient PV plants. The paper proposes a resampling method, so-called bootstrap, able to approximately describe the sampling distribution. Its peculiarity consists in taking information about the population of the data (whichever its distribution is) and giving quickly preliminary information about the operation of the PV plants.**

## I. INTRODUCTION

The dependence of the system response from many extrinsic factors, such as insolation intensity, ambient temperature, cell temperature, air velocity, humidity, cloudiness and pollution have to be taken into account in the design of a PV plant as well as during its operation and several models to evaluate the effects of such uncertainties [1-3] and the critical choice of the electrical components [4]-[5] have been proposed.

Well-timed selection of mis-operating sub-systems and out-of-service prevention are important challenges for reliable operation of any solar energy system. For this aim, standard benchmarks [6], called “Final 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. Unfortunately, they shows drawbacks and a two-step procedure, based on descriptive and inferential statistics, has been proposed [7]. The procedure, based on the energy data stored in the data-logger of PV plant, has been implemented in Matlab environment.

Even if energy data of a whole year are not available, it is possible to extract information about the population applying the procedure described in [8], while [9]-[10] propose a user-friendly Labview interface for an easy application of both the procedures.

When government incentives are available, the PV market becomes very attractive for new PV plants installations as well as for Operation and Maintenance (O&M) enterprises. Several O&M enterprises have a lot of PV plants to be monitored and a data-logger on each PV plant. An analysis based on the whole population of the energy data (the economic incentives are often based on the energy production!) is possible if few PV plants with small rated power have to be monitored. But, when several PV plants have to be monitored and each of them has great rated power ( > 1 MWp), the amount of the data explodes and the processing time grows exponentially. In these situations we are in presence of Big Data (BD), defined as “*high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization*” [11].

The possibility to acquire, to manage and to process BG is rather recent, thanks to the establishing of the systems for hyperscale data storage (able to expand rapidly and effectively), to the databases and platforms for data analysis, to the increased computing power of the PCs, to the introduction of high performance computing, which allows the parallel processing of data in order to support quickly, efficiently and reliably the complex application programs. This technological aspects have allowed an evolution of the statistical tools in several daily activities, over of all related to the social web applications. In fact, in those fields both the classical and the Bayesian statistics have already been substituted by the Exploratory Data Analysis (EDA), promoted by Tukey [12], modifying the approach to the data analysis (Fig. 1).

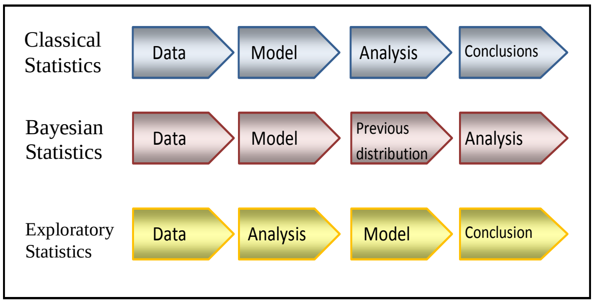

*Fig. 1 Different statistical apporaches*

The traditional approaches build the model and, successively, do the analysis, whereas the EDA does the analysis and then builds the model. Some computer applications use the *stream analysis algorithms* to manage BD. A data stream can be defined as a temporal

ISBN-14: 978-92-990073-2-7

ordinate sequence of values. In recent years, the management of these information flows has been based on Data Stream Management Systems (DSMSs) [13]. A DSMS can be considered as an extension of a classic DataBase Management System (DBMS), able to query continuously sequence of sorted data. In fact, while the traditional DBMSs are designed to work on mostly static data, where updates are infrequent, the DSMSs are specialized in working with highly variable data that are continuously updated. Similarly, queries made through a DBMS are performed one time to return a definite answer, while queries on DSMS wonder cyclically data to provide updated answers as soon as new information becomes available. In presence of BD and diversity of information, DSMSs integrate advanced techniques for processing streams that simplify the queries, but obtaining approximate results. Nevertheless, this new approach is not still used for real time monitoring of several PV plants by a unique centre, then the traditional approach based on sampled data can be used for a rough preliminary analysis and the identification of bad-working PV plants.

Among the several re-sampling techniques, this paper proposes the application of the bootstrap [14]. This technique is independent from the population distribution; then it can be directly applied to the available data without any verification.

The aim of this paper is to show that the sampling distribution can be substituted by the bootstrap distribution preserving the information and speeding up the processing time.

The paper is organized as follows: Section II shows the number of data needed to the monitoring, Section III presents the general properties of the bootstrap technique, while results on a real PV plant and conclusions are proposed in Section IV and Section V, respectively.

## II. MONITORING CENTER OF PV PLANTS

A PV plant with 1 MWp rated power can be constituted by about 250 strings. If the data-logger samples each 10 minutes for 10 hour/day, 6 data set/hour and 60 data set /day for each string have to be stored; then, 15.000 data set /day for 250 strings and 5.475.000 data set/year for the whole PV plant. Each data set is constituted by several measures (voltage (DC-AC), current (DC-AC), power (DC-AC), energy (AC), cell temperature, environment temperature, radiance, ecc.). Even if an analysis is based on only 4 variables as three-phase voltage and current, energy, cell temperature and radiance, 54.750.000 data have to be processed. The data to be processed from an O&M enterprise which has to monitor, for example, 10 PV plants with 1MWp rated power, are about 545.750.000!

Because the data of different plants are independent each other, the actual trend is the adoption of an informatics structure consisting of a control room with a server box linked to the data-logger of each PV plant (from the monitoring point of view the PV plants constitute a ”constellation” of plants connected to the same server), as shown in Fig. 2. Even if the server box can manage sequentially the data of different plants, the processing time is high with respect to a quick and frequent information about the operation of each PV plant in order to prevent failures and lack of the incentives.

This paper proposes to carry out preliminary analysis on each PV plant utilizing sampled data in role of the whole data population, in order to manage a smaller dataset for each PV plant. If an anomaly is revealed on a specific PV plant, a successive in depth analysis (based on the whole data population of that PV plant) will be carried out. In the following, all the considerations about the sampling will regard only one PV plant, being intended that the concepts have to be applied to each PV plant.

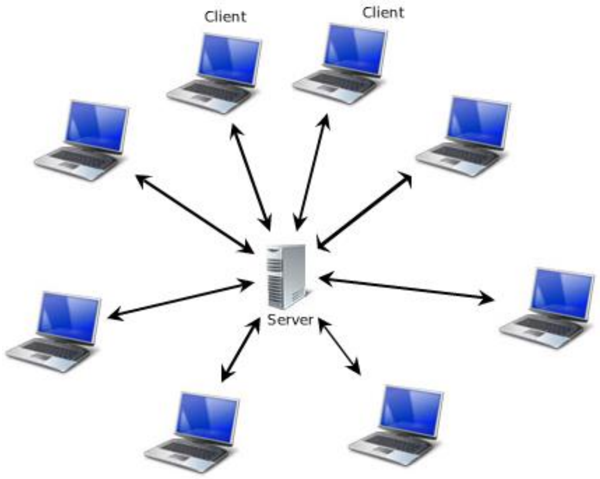

*Fig. 2. Monitoring system of a constellation of PV plants.*

The sampling gives correct information if the sample is *representative* of the population. Sometimes, in order to obtain a representative sample, it is needed to consider more and more samplings and the final sample is obtained after processing the previous ones. Obviously, if the population is numerous, the total sampling time can be high for a prefixed significance level and the sampling advantage is lost, while the inaccuracy remains.

In the next Sections it will be shown that the bootstrap technique: a) allows to define a *representative* sample of the energy data stored in the data-logger of a PV plant; b) is more efficient than ordinary sampling technique, requiring a smaller computational time.

For the former issue, it will be shown that the bootstrapped sample matches the properties of the distribution usually obtained by means of many samplings. For the latter one, a comparison between the computational times related to the bootstrap technique and the ordinary sampling technique will be performed.

## III. BOOTSTRAP TECHNIQUE

This section deals with: a) the differences between bootstrap technique and ordinary sampling; b) the three characteristic properties of any distribution, useful for comparing the results of the next section; c) the two random sources of variation between the population statistics and the bootstrap ones.

### *A. Differences between bootstrap and sampling*

Statistical inference is based on many random samples from the population with the aim to find the sampling distributions of sample statistics (mean, variance, etc.). Instead, the bootstrap is a resampling technique (as jack-knife and permutation test are) able to approximately find the sampling distribution from just one single sample [14]. In fact, in place of many samples from the population, the bootstrap creates many resamples by repeatedly sampling with replacement from only one random sample. Each resample is the same size as the original random sample.

Then the bootstrap distribution gives information about the sampling distribution, while the original sample represents the population from which it was drawn. The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic, based on many samples [14].

### *B. Bootstrap properties*

There are three properties which characterize any distribution: shape, center and spread. Let us analyze these properties for a bootstrap distribution.

- Shape: the central limit theorem says that the sampling distribution of the sample mean $\overline{x}$ is approximately normal if the sampling number *n* is large. The bootstrap distribution is nearly normal, then its shape is close to the shape we expect the sampling distribution has.
- Center: we know that the sampling distribution of a sample mean $\overline{x}$ is centred at the population mean µ, i.e. $\overline{x}$ is an *unbiased* estimate of µ. The bootstrap distribution is centred at the mean of the original sample, but this last one is biased with respect to the population mean. Then the mean of the bootstrap distribution has little bias as an estimator of the population mean. So, the resampling distribution behaves with respect to the original sample as the sampling distribution behaves with respect to the population.
- Spread: It is the variation among the resample means and is said Bootstrap Standard Error (BSE) of $\overline{x}$. It represents the standard deviation of the bootstrap distribution of $\overline{x}$.

We found that, as shown in [14], the bootstrap distribution created by resampling matches the properties of the sampling distribution.

### *C. Variation of the bootstrap statistics*

Finally, when bootstrap technique is used, two sources of random variation between the population statistics and the bootstrap ones have to be taken into account:

- variation due to the original sample;
- variation due to the bootstrap resamples from the original sample.

Let us consider, for example, the variation introduced on the mean value. The mean of the original sample has a bias with respect to the population mean; the mean of each successive resample, based on the first original sample, has a bias with respect to the mean of the original sample and then with respect to the population mean. The inaccuracy of the mean value $\overline{x}$ of the starting sample affect all resamplings, while the random resamplings add a little variation to $\overline{x}$. Then, a limited number of resamplings is needed to obtain a significant sample.

## IV. RESULTS AND DISCUSSION

As explained in previous section, the bootstrap technique has to be applied to each PV plant belonging to a constellation, in order to have smaller data sets to be quickly analyzed by means of statistical approaches.

For this aim, it is necessary that: a) the bootstrapped sample matches the sampling distribution (i.e. the bootstrap sample has to be representative of the population); b) the computational time has to be small. In other words, it is necessary to show that the sampling distribution can be substituted by the bootstrap distribution preserving the information and speeding up the processing time.

To test the efficiency of the bootstrap technique, in this section the methodology will be applied to the energy dataset of a real PV plant (about 20 kWp rated power) located in the south of Italy. Obviously the results are similar, whichever peak power is considered; it is different only the amount of operations and the related computational time.

The PV plant under test (19,8 kWp rated power) consists of 6 sub-arrays composed of two parallel connected strings, each of them having 11 photovoltaic 150 Wp modules. The total power amount for a single array is 3300 Wp. For each sub-array a 3000 W inverter has been installed, after considering the total losses (mismatching, voltage drop due to the wires, etc.).

The whole data population is constituted by the values of the energy produced in the year 2007 by the PV plant. One value of energy for each sub-array is stored in the datalogger each 10-15 minutes; then, 23.795 energy values for each sub-array are available for the whole year. Fig. 3 reports the population distribution for each sub-array and it highlights that the data are not normally distributed; the y-axis of each inverter reports how many times an energy value of x-axis compares in the respective data set.

Fig. 4 represents the sampling distribution of the sample mean. In this figure, the sample size for each sub-array is 20% of the data population. It can be seen the validity of the central limit theorem, for which, as n increases, the sampling distribution of the sample mean tends to be normally distributed around the population mean μ, even if the population is not normally distributed.

Moreover, μ and $\bar{x}$ (sample mean) are almost equal because the sample mean is an unbiased statistic; obviously, if a greater sample size is chosen, a smaller difference between μ and $\bar{x}$ is expected.

Fig. 5 reports the bootstrap distribution based on a single sample for each sub-array whose size is 20% of the population data (as in the case of the sampling distribution).

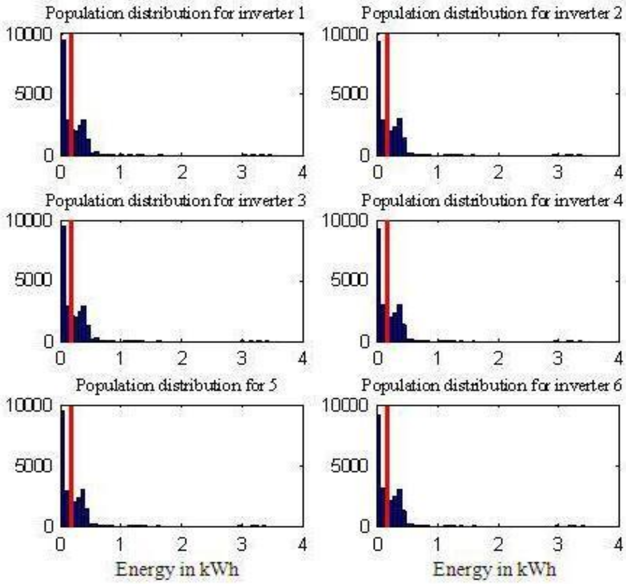

*Fig. 3. Population distribution of each sub-array.*

Several resamplings with replacement are done from this unique sample. As the mean of one sample is usually biased with respect to the population mean (red), it results that the bootstrap distribution of the bootstrap mean (fuchsia) is unbiased respect to the sample mean (blue) but is biased with respect to the population mean. Tab. I reports, for each one of the six inverters, the values of the means of the populations, the samples, the bootstrap re-samples and the related BSE. Moreover, the resamplings for the bootstrap are done in a shorter time than the sampling one (0.61s versus 12.39s); it depends from the possibility to choose the number of resamplings. For this specific case (but the same results have been obtained for a lot of repetitions), the bootstrap time is about 5% of the sampling one.

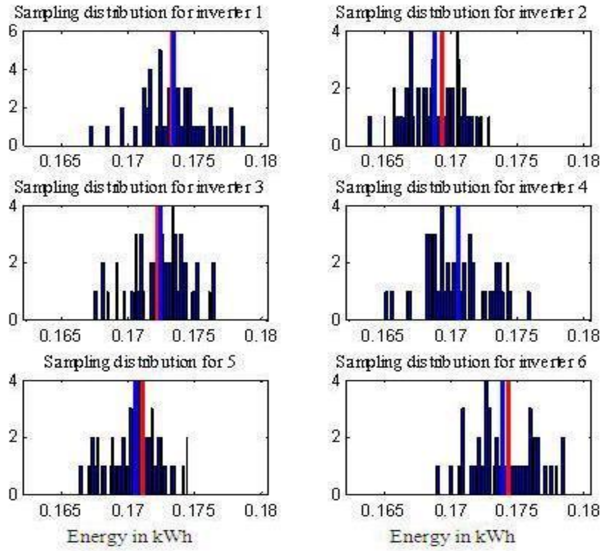

*Fig. 4. Sampling distribution of the mean of each sub-array.*

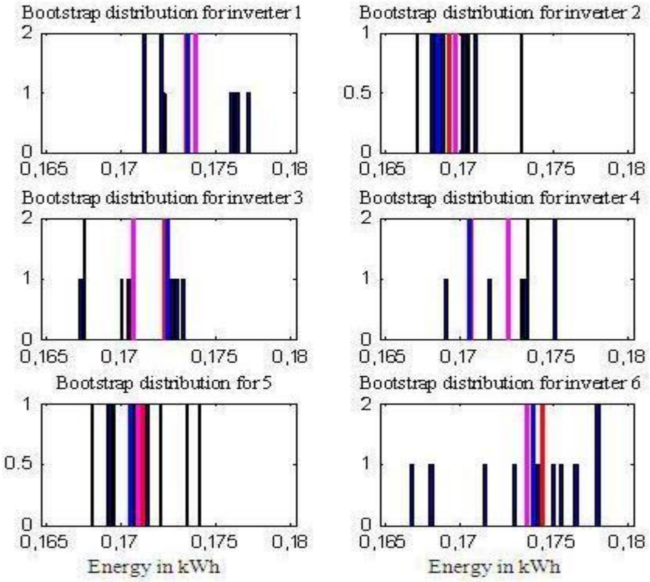

*Fig. 5 Bootstrap distribution of bootstrap mean for sub-arrays*

*Table. I. Energy values (in kWh) of the means for populations, samples, bootstrap resamples and BSE*

| | Inv 1 | Inv. 2 | Inv. 3 | Inv. 4 | Inv. 5 | Inv. 6 |
|---|---|---|---|---|---|---|
| $\mu$ | 0.1734 | 0.1694 | 0.1722 | 0.1706 | 0.1711 | 0.1744 |
| $\bar{x}$ | 0.1734 | 0.1688 | 0.1724 | 0.1705 | 0.1705 | 0.1739 |
| $\bar{x}*$ | 0.1739 | 0.1698 | 0.1706 | 0.1726 | 0.1709 | 0.1736 |
| BSE | 0.0024 | 0.0016 | 0.0022 | 0.0020 | 0.0019 | 0.0025 |

### A. *Variation due to the resamplings based on the original sample*

Fig. 6 reports other bootstrap distributions based on the same samples in order to show that the random original sample of each inverter is responsible for the most variation with respect to the populations, as explained in Sec. III. In fact different resamplings based on the same six original samples produce only little variation with respect to that due to the random original sample. Tab. II reports the numerical values of the same variables of Tab. I. Obviously, the values of the rows 1 and 2 are equal while the values of the third and the fourth row (related to the resamplings) are different.

It can be noted that the differences are very small, as expected. Once again the resamplings for the bootstrap are done in a shorter time than the sampling one (0.60s versus 12.39s).

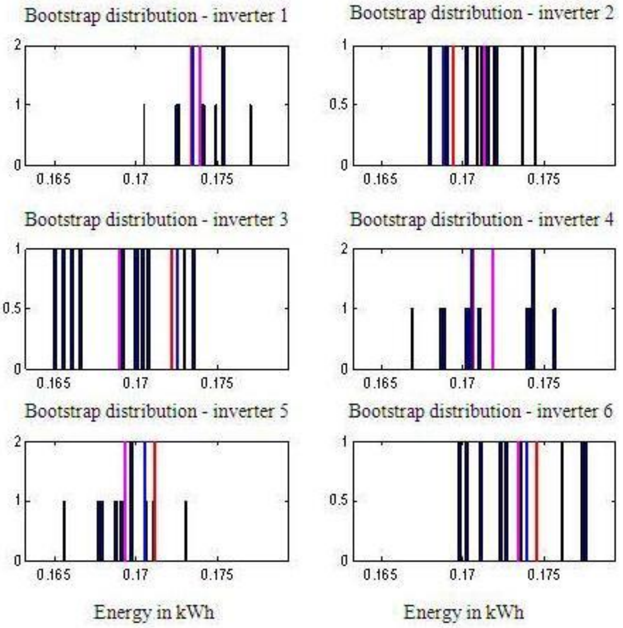

*Fig. 6. 2nd Bootstrap distribution of the bootstrap mean for each sub-array.*

*Table. II. Energy values (in kWh) of the means for populations, samples, bootstrap resamples and related BSE of the 2nd case.*

| | Inv 1 | Inv. 2 | Inv. 3 | Inv. 4 | Inv. 5 | Inv. 6 |
|---|---|---|---|---|---|---|
| $\mu$ | 0.1734 | 0.1694 | 0.1722 | 0.1706 | 0.1711 | 0.1744 |
| $\bar{x}$ | 0.1734 | 0.1688 | 0.1724 | 0.1705 | 0.1705 | 0.1739 |
| $\bar{x}*$ | 0.1739 | 0.1713 | 0.1690 | 0.1718 | 0.1693 | 0.1734 |
| BSE | 0.0016 | 0.0020 | 0.0011 | 0.0027 | 0.0022 | 0.0023 |

### B. *Variation due to the original sample.*

This sub-section reports the results of the bootstrapping on the basis of a new random original sample in order to highlight that the variation due to the original sample is stronger than that due to the random resamplings. Fig. 7 reports the second set of samplings based on the same population data, while Fig. 8 reports the bootstrapped mean distribution for each sub-array. The sampling distribution as well as the bootstrap distribution give similar values as the previous ones and then the mismatches are limited in a small range. Once again, the resamplings for the bootstrap are done in a shorter time than the sampling one (0.77s versus 12.39s).

The obtained results highlight that when BD have to be processed for monitoring the PV plants in a constellation, the bootstrap technique can be an effective approach. It has two main advantages. The former one is that it does not depend on the dataset distribution, then the classification of the distribution (usually required from other sampling techniques, implying processings or analyses) is not needed.

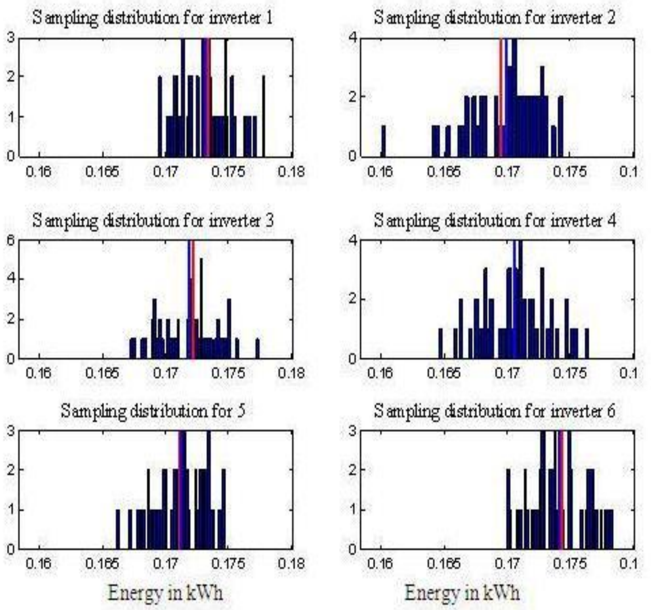

*Fig. 7. New sampling distribution of the mean value for each sub-array.*

The latter one is that it is a time-saving procedure with respect to simple sampling techniques. The main advantage for enterprises of O&M to use this approach to monitor PV systems belonging to a constellation is the ability to make a quick preliminary analysis on the state of health of all the PV plants, in order to identify the only PV plants with anomalies. Then, an in depth analysis can be done only for these critical PV plants. This approach allows you to focus your time and attention only on PV

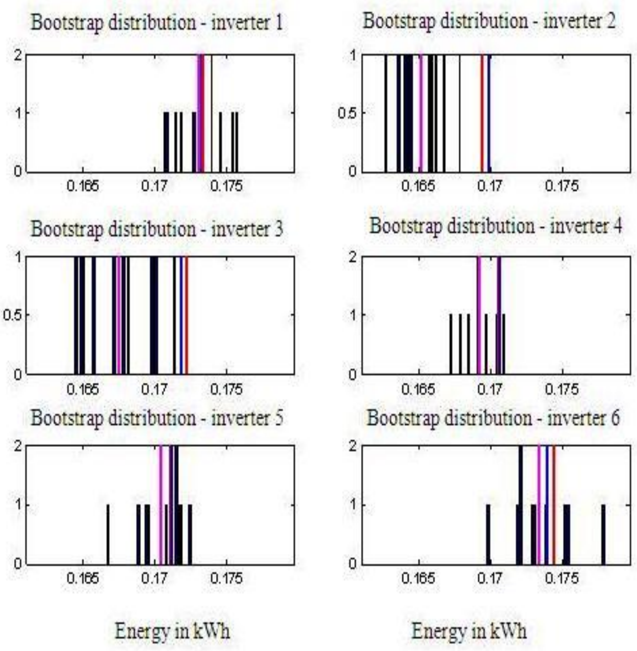

*Fig. 8. Bootstrap distribution of the bootstrap mean for each sub-array on the basis of the new samples.*

systems with anomalies.

## V. CONCLUSIONS

The monitoring of energy efficiency of large PV plants requires long time for the statistical treatment of the acquired data. The problem becomes harder when a lot of large PV plants have to be monitored from a unique supervision center. To reduce the processing time it is possible to represent the population of data by means of correct samplings which allow to perform a faster evaluation of the operating condition of plants, postponing the analysis of the full data set only if anomalies are found in the first analysis. Nevertheless the generation of a significant sample by means of statistical inference needs a consistent computational time. To reduce the sampling time, the paper proposes the bootstrap methodology.

## VI. ACKNOWLEDGEMENTS

The paper has been written within the framework of the project “RES NOVAE - Reti, Edifici, Strade - Nuovi Obiettivi Virtuosi per l’Ambiente e l’Energia”. This project is supported by the Italian university and Research National Ministry and competitiveness program that Italy is developing to promote “Smart Cities Communities and Social Innovation.

## REFERENCES

[1] Edmundo Romàno, Ricardo Alonso, Pedro Ibanez et alt. "Intelligent PV module for grid-connected PV systems", in IEEE Trans. On Industrial Electronics , vol. 53, No. 3, Aug. 2006, pp. 1066-1073.

[2] Il-Song Kim, Myung-Bok Kim, Myung-Joong Youn, "New maximum power point tracker using sliding-mode observer for estimation of solar array current in the grid- connected photovoltaic system", in IEEE Trans. On Industrial Electronics , vol. 53, No. 4, Aug. 2006, pp. 1027-1035.

[3] Weidong Xiao, Magnus G.J. Lind, William G. Dunford, Antoine Capel, "Real-time identification of optimal operating points in photovoltaic power systems", in IEEE Trans. On Industrial Electronics , vol. 53, No. 4, Aug. 2006, pp. 1017-1026.

[4] McSharry P.E., "Assessing photovoltaic performance using local linear quantile regression", in Proceeding Energy and Power System 2006, ACTA Press.

[5] Ziyad M, Salameh, Bogdan S. Borowy, Atia R. A. Amin, "Photovoltaic module- site matching based on the capacity factors", in IEEE Trans. On Energy Conversion , vol. 10, No. 2, June 1995, pp. 326-332.

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

[7] 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, 2009, pp. 4456-4464.

[8] S. Vergura, G. Acciani, V. Amoruso, G. Patrono, "Inferential Statistics for Monitoring and Fault Forecasting of PV Plants", IEEE-ISIE'08 - International Symposium on Industrial Electronics (ISBN 978-1-4244-1666-0), June 30 to July 2, 2008, Cambridge, UK, pp 2414-2419.

[9] S. Vergura, E. Natangelo, "Labview Interface for Data Analysis of PV Plants", IEEE-ICCEP 2009, 9-11 June, 2009, Capri, Italy, pp. 236-241.

[10] S. Vergura, E. Natangelo "Labview-Matlab integration for analyzing energy data of PV plants", ICREPQ 2010, March, 23-25, 2010, Granada, Spain (ISBN 978-84-613-7543-1).

[11] Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner, June 2012.

[12] J.W. Tukey, “The Future of Data Analysis”, Annals of Mathematical Statistics, n. 33, pp. 1-67, 1962.

[13] L. Golab, M.T. Özsu, “Issues in data stream management”, ACM SIGMOD Record, Vol. 32, Num. 2, pag. 5-24, 2003.

[14] Efron, B., "Bootstrap Methods: Another Look at the Jackknife". The Annals of Statistics, 7 (1), pp 1-26