2017 International Conference on Big Data, IoT and Data Science (BID)
Vishwakarma Institute of Technology, Pune, Dec 20-22, 2017

Roadmap to Prepare Distribution Grid-Tied Photovoltaic Site Data for Performance Monitoring

Aditya Sundararajan
Department of Electrical and Computer Engineering
Florida International University
Miami, USA
asund005@fiu.edu

Arif I. Sarwat
Department of Electrical and Computer Engineering
Florida International University
Miami, USA
asarwat@fiu.edu

***Abstract*—One of the key analytics conducted on a gridtied Photovoltaic (PV) system is the periodic monitoring of its performance. It is expected that with increased PV penetration into the distribution smart grid in the future, quality and integrity of the data required to conduct such analytics will be crucial. While data processing and management tools for smart grid in the literature use cloud, distributed file management and parallel processing, the latency and computation requirements specific to performance monitoring need more lightweight, descriptive methods. This paper provides a systematic roadmap to analyze data collected from a real distribution grid-tied 1.4*MW* PV power plant for completeness, consistency and integrity, with the objective of using it for performance monitoring. To ensure the data's integrity is not compromised, the distribution of processed data is compared with that of the raw data. This paper makes one of the first few attempts to provide a comprehensive approach for data scientists to clean and prepare grid-tied PV data for site-level performance monitoring.**

***Index Terms*—smart grid, PV big data, performance monitoring, data processing**

I. Introduction

With increasing levels of solar photovoltaic (PV) integration into the smart grid's distribution network, applications such as performance monitoring have been gaining importance. The conceptual layout of smart grid shown in Fig. 1 provides the context for grid-tied PV sites. It can be seen that the grid comprises five key power system layers (Power Layer), with different *Modules* within each. Each Power Layer has a corresponding *Computation Layer* comprising Internet of Things (IoT) sensor networks that ubiquitously collect, bidirectionally communicate, and process data [1]–[3]. These layers constantly interact with a central intelligence, usually located at the Command and Control Center (CCC), where bulk of information ingestion and processing takes place.

The intelligence associated with the Power and Computation Layers are currently centralized, with IoT sensors constantly measuring and sending data to the central processing engines that explore, mine, visualize, clean and ready data for power system applications. However, with emerging shift in computing from centralized to distributed architecture, as advocated by the General Electric (GE) through its EdgeComputing.

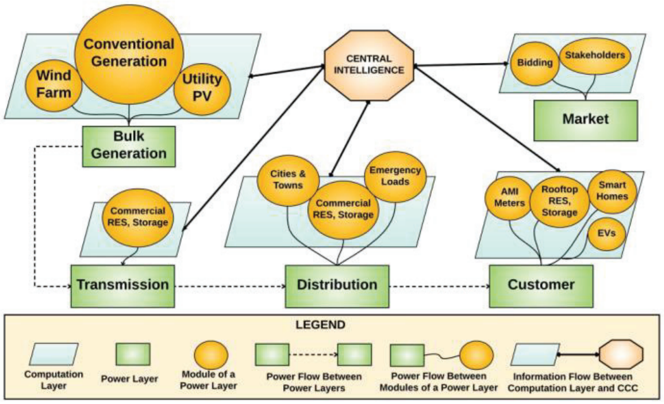


Fig. 1. Conceptual layout of smart grid.

Platform and OpenFog Consortium through its Fog Computing Platform, lightweight processing (descriptive and diagnostic methods like cleaning, structuring, quality-checking, and missing data handling) is expected to be done at local PV sites [4]. Grouped under "site-level analytics", they form the prime focus of this paper. This requirement in further bolstered by the fact that there is little to no visibility and control for the operators at CCC over remotely located PV sites [5]–[7].

Hence, electric utility employees such as site owners, installers, operators and integrators rely on different performance monitoring software that collect, process and visualize data and performance characteristics for them [8]. However, most of these software still employ centralized framework such as cloud-based analytics, which negate the emerging distributed computing paradigm. A few software integrate their analytics with Apache Hadoop Spark or Storm which support distributed stream processing [9]–[12]. However, a site-level analytic like performance monitoring does not need data of that high resolution either. Hence, a middle ground between cloud and stream processing is required, which the existing research does not appropriately address.

The key contribution of this paper is a systematic roadmap to process the multivariate time-series data collected from a real

The work for this paper has been sponsored by the National Science Foundation (NSF) Grant No. CNS-1553494.

978-1-5090-6593-6/17/$31.00 ©2017 IEEE

1.4*MW* distribution grid-tied PV site to be used for its performance monitoring. This roadmap serves as a useful guideline for utility employees to simplify their current methodologies for site-level analytics. The paper demonstrates tested descriptive and diagnostic methods like correlation, multivariate imputation and Maximum Likelihood Estimation (MLE) by building and executing scripts written in R and utilizing one year's data from a real PV site. The processed data is then compared with the original raw data to ensure its integrity has not been compromised. Once fully developed and integrated with the utilities' business models, the roadmap would not only be useful for performance monitoring, but also for other site-level analytics such as component-level reliability analysis, PV generation forecasting, and economic dispatch, since all of them need good quality data.

The rest of the paper is organized as follows. Section II briefly discusses the different data analysis approaches that exist in the literature, highlights their shortcomings, and signifies the need for the proposed roadmap. Section III introduces and describes the roadmap and its constituent steps, demonstrating them with two specific datasets: PV Module Temperature and Ambient Temperature. Finally, Section IV documents the conclusion of the proposed effort and future work.

II. Related Work

The challenge of big data in smart grid is not fully explored in the literature. Most existing works focus on addressing big data challenges as a single problem for the entire grid, when in reality, computational resources and latency needs become more stringent towards the grid's edge than at the center [13]. Emerging paradigms such as Edge and Fog Computing propose to move lightweight (descriptive and diagnostic) analytics towards the grid's edge where PV sites are also located, and keep heavyweight (predictive and prescriptive) analytics at CCC. It is, hence, important to develop an approach that is cognizant of decentralizing intelligence.

Different methods have been proposed to process smart grid data. At a broader level, these approaches involve the cloud, parallel processing, and Distributed File Management System (DFS) platforms like Apache Hadoop. The authors in [14] emphasize the significance of big data and its analysis for smart grid, and also propose an architecture for consumer analytics which begins with lightweight analyses to heavyweight ones. Hadoop MapReduce, Storm and Spark are proposed for batch, stream and hybrid processing and management [9], [10]. While Hadoop technologies are capable of effectively handling the increasing proportion of available and emerging data, they also require significant computation power which cannot be expected to be available at PV sites where resources are constrained.

Distributed stream computing platforms like Apache Spark and Storm operate on ”windows” of data, performing lightweight analytics on windowed chunks before working on aggregated windows for heavier computation. The Spark and Storm platforms are advantageous for site-level analytics since they can work with limited resources and are latency-aware. However, stream computing is not a paradigm that best fits the requirement of performance monitoring since it uses data aggregated over 15-minute intervals. A few other works recommend data management and processing, but their scope extends across the entirety of smart grid and their effectiveness for site-level analytics might not be significant considering specific resource constraints and latency needs at local level [15], [16].

III. The Proposed Roadmap

The proposed roadmap, specific for site-level analytics, is shown in Fig. 2. It comprises five key steps: 1) Identify the objective of data processing; in this case, it is site performance monitoring, 2) Collect raw data required to meet the objective; in this case it is data on site inverters and weather, 3) Estimate the missing data using a multivariate imputation method, 4) Explore the data for its properties using correlation and MLE, and 5) Compare the processed data with original, raw data statistically using a Cullen and Frey Graph. After this step, the data is considered ready to be used for site performance monitoring. For the purposes of brevity, this paper demonstrates the roadmap specifically for two attributes: PV module temperature and ambienttemperature. It is to be understood that the same sequence of steps can be applied to other attributes as well.

A. Identify the Objective

The power generated by PV site depends on different factors like: a) Irradiance: intensity of solar radiation falling over the PV module, its angle of incidence, and the tilt-angle of the module; b) Weather: PV module temperature, ambient temperature, precipitation, humidity and wind velocity; c) Inverter: yield, service downtime, efficiency, and power factor; d) System losses due to soiling, dirt, and DC-AC conversion, modules' temperature coefficient, and many more. Due to the influence of so many factors, PV site performance varies with time, seasons and nameplate capacity [17]–[19]. Accordingly, there are different metrics such as yield, Performance Ratio (PR), Energy Performance Index (EPI), Power Performance Index (PPI), and capacity factor.

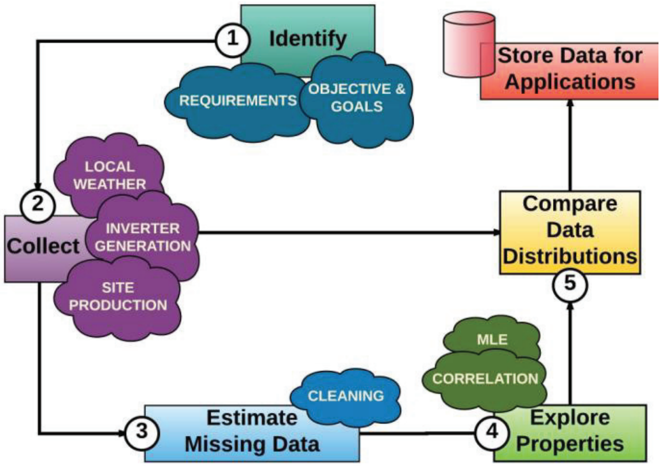


Fig. 2. Flowchart of the proposed roadmap.

Each of these metrics require estimated generation of the site, which is calculated using the following equation [20]:

$$kWAC_{estimated} = kWDC_{rated}\frac{Irr(t)}{1000}[1 + \frac{Temp\_Coef}{100}\{T(t) - 25\}]P \quad (1)$$

Here, $kWAC_{estimated}$ is the estimated power output from a PV site that has a rated power of $kWDC_{rated}$. $Irr(t)$ is the average solar irradiance (in kW/m^2) falling over the entirety of the site's PV modules at a given time*t*, and *Temp Coef*is the percentage temperature coefficient of the PV modules. While $T(t)$ is the module temperature, P is the cumulative multiplicative value of different factors like module mismatch, dirt derate, losses due to cabling, and losses due to DC-AC conversion.

B. Collect the Required Raw Data from the Site

A layout of the 1.4*MW* PV site considered for this study is shown in Fig. 3. The site comprises a total of 4,480 modules organized into four arrays, with 224 parallel strings of 20 serially connected modules each. 46 string inverters are connected to these modules, with 40 of them tied to 100 modules each, and 6 of them to 80 modules each. These inverters convert the DC power generated by modules into AC power, and a combiner panel aggregates the converted power. A revenue-grade production meter installed on-site records net energy produced by the site. A pad-mounted liquid power transformer is located at the Point of Common Coupling (PCC) where the site connects to the distribution feeder owned and operated by the utility.

An on-site Data Acquisition System (DAS) exists, which records the readings every three minutes by polling device registers allocated for the field inverters, on-site weather station, and meter. It then sends this data over a secure Global System for Mobile (GSM) channel to its cloud every 15 minutes. For illustrating the roadmap, site data was collected for a period of one year (August 1, 2016-July 31, 2017). This amounts to 35,040 observations.

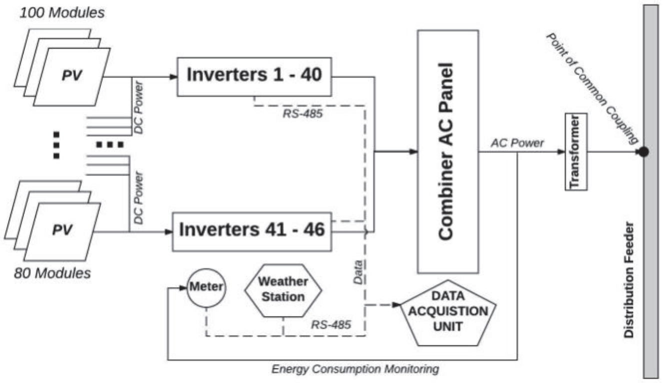


Fig. 3. Layout of the 1.4*MW* PV Plant.

C. Estimate Missing Data

An incomplete dataset is one of the key reasons for poor quality. There exist many ways to deal with missing values. The simplest one is to ignore the records with such values or substitute them with zeros. However, if a dataset has more than 5% missing values, ignoring or zeroing an entire record implies the loss of useful information. Some other ways include replacing with column mean, column median, or average of neighboring values. Machine learning algorithms like Support Vector Machines (SVM) or Neural Networks have also been proposed to estimate missing values. However, for site-level analytics, there are circumstances when not enough historical data are available to train a learning algorithm, or to validate the model. To this effect, the paper considers using Fully Conditional Specification (FCS) to estimate missing data by decomposing an *n*-dimensional problem into *n* one-dimensional sub-problems [21]. This method is practically more efficient since it ensures the imputed dataset retains its structure and shape, and does not compromise its logical consistency [22]. Although computationally more intensive than other imputation methods, FCS offers an acceptable tradeoff between complexity and structure-preservation.

There exist significant portion of energy data that was missing, due to causes such as inverter device failure and communication delay. There were missing values even in module temperature and irradiance due to communication failure, but they were ignored for this study as the missing values amounted to less than 1% of the dataset.

Missingness in a dataset can be of two types: 1) Missing Completely at Random (MCAR), or 2) Missing Not at Random (MNAR). MCAR datasets are easier to impute since the missing values do not have a dependency with other missing or present values. However, MNAR needs careful scrutiny, as the missing values have a relationship between them and with present values.

Hence, it is important to first test the data to ensure if it is MCAR or MNAR, and then estimate accordingly. Little's MCAR test on the data provides a chisquared significance test under the null hypothesis H_0 that the data has MCAR type missing values [23]. If the hypothesis is rejected, it implies the data has MNAR type values. For $\alpha = 0.05$, the test revealed that H_0 cannot be rejected, and hence, the data is most likely MCAR. In such a case, a multivariate imputation employing FCS can be applied.

The process imputes missing values by iterating over conditional densities and preserves the relationship between variables and the uncertainty in these relationships, the data's distribution, and works on the dataset one column at a time using Gibbs sampling, as shown in Fig. 4. Being multivariate imputation, the attributes having missing values are treated as targets and their values are imputed as conditional densities given the values in other attributes.

D. Explore the Data for its Properties

With the data cleaned of missing values, one of the first steps typically conducted is a correlation study to understand different attributes. Shown in Fig. 5, the correlation matrix depicts the coefficient of correlation, R, for different pairs of attributes. The scale to the right shows the color code where R is bounded by the region $[-1,1]$. It can be seen that R is mostly positive, implying that these variables share a direct proportionality relationship. Specifically, it can be seen that irradiance, module temperature and ambient temperature have a significant correlation with site's power recorded by the production meter. While irradiance has $R = 0.94$, module and ambient temperatures have $R = 0.4$ and $R = 0.2$. This, however, does not determine the dependency between the variables themselves. For instance, it cannot be concluded that high irradiance implies greater PV generation. It simply states that an increase in irradiance is seen alongside an increase in PV generation.

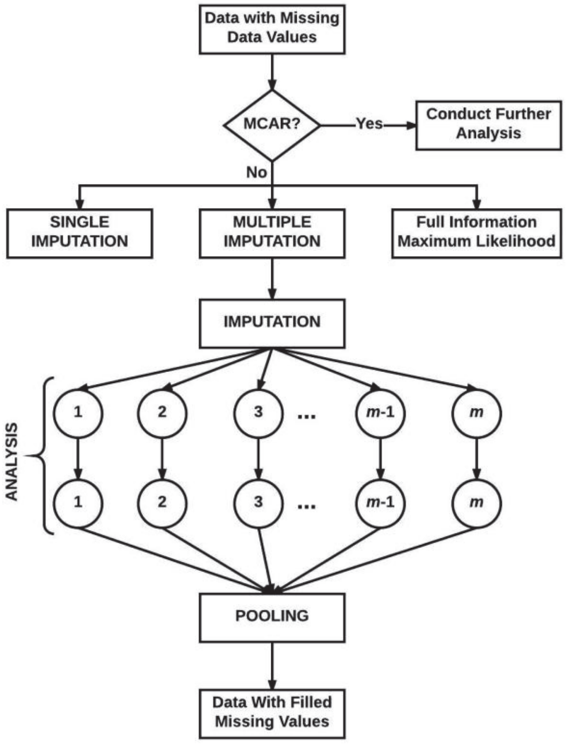


Fig. 4. Flowchart for missing data estimation using multivariate imputation.

Maximum Likelihood Estimation (MLE) is used to estimate the parameters of a sample set of data considering they fit a model, so as to maximize the likelihood of obtaining these data values given the estimated parameters [24], [25]. Given a random sample, $X_1,X_2,...X_n$with a sample of observed values, $x_1,x_2,...,x_n$whose Probability Density Function (PDF) is determined by an unknown parameter θ, the objective is to determine an estimator $M(X_1,X_2,...,X_n)$ such that $M(x_1,x_2,...,x_n)$ is a good estimate of θ. The joint PDF of $X_1,X_2,...,X_n$, called $L(\theta)$ can be defined as $L(\theta) = P(X_1 = x_1,X_2 = x_2,...,X_n= x_n) = f(x_1;\theta) \cdot f(x_2;\theta)...f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, where $f(x_i,\theta)$ is defined as the PDF of each X_i, $i\in [1,n]$. MLE is a powerful technique to estimate the goodness of fit of data and has the following key asymptotic

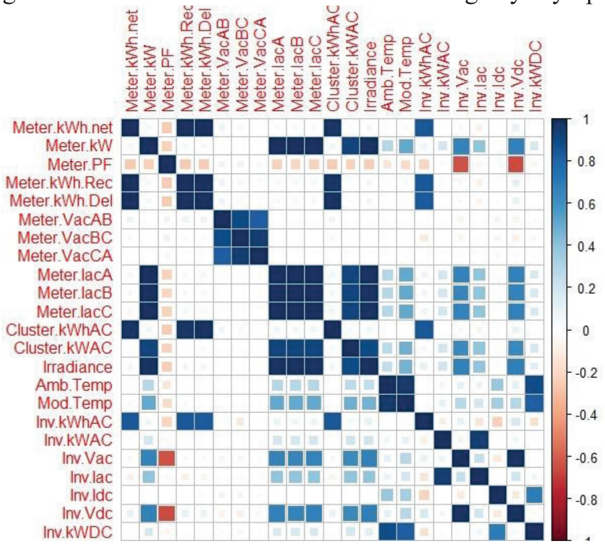


Fig. 5. Correlation of different PV site parameters.

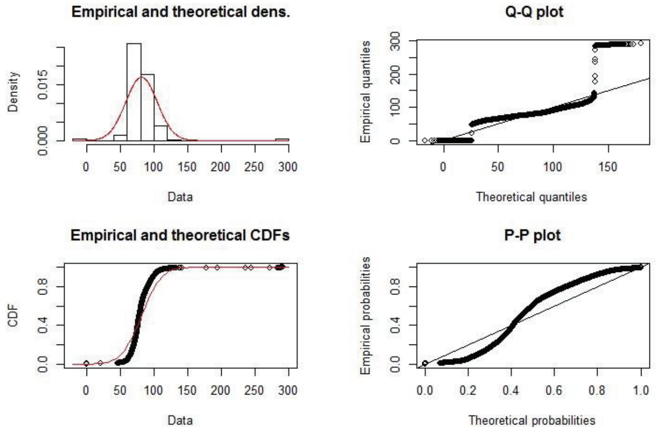


Fig. 6. Fitting PV Module Temperature using MLE.

properties that make it a significant approach in the roadmap: the ML estimator, $\hat{\theta}$, is consistent in that the sequence of MLE probabilities converges to estimated value; asymptotic efficiency and normality [26]–[28]. Further, their tendency to bias decreases as the sample size grows, and the likelihood functions could be used for statistical hypothesis testing. The results from applying MLE on PV module temperature and ambient temperature datasets are illustrated by Fig. 6 and 7, respectively.

The Quantile-Quantile (Q-Q) and Probability-Probability (P-P) plots for PV module temperature shown in the top and bottom right of Fig. 6 illustrate the agreement between the theoretical and empirical quantiles and probabilities, respectively. The empirical data is drawn from the sample observations while the theoretical data belongs to the known beta distribution. While Q-Q plot compares the quantiles, P-P plot compares the Cumulative Distribution Functions (CDFs), which is affirmed by the agreement between the theoretical and empirical CDFs shown in the bottom-left of the same figure. It can be concluded both mathematically as well statistically that beta distribution can be viewed as the most likely fit for PV module temperature data. A similar analysis on ambient temperature shown in Fig. 7 revealed that normal distribution fits it best. This analysis can be conducted on other collected datasets as well.

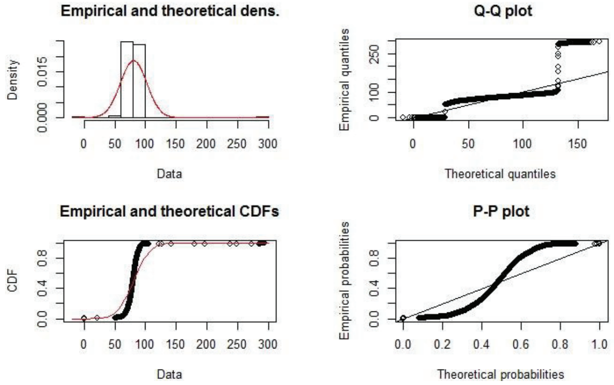
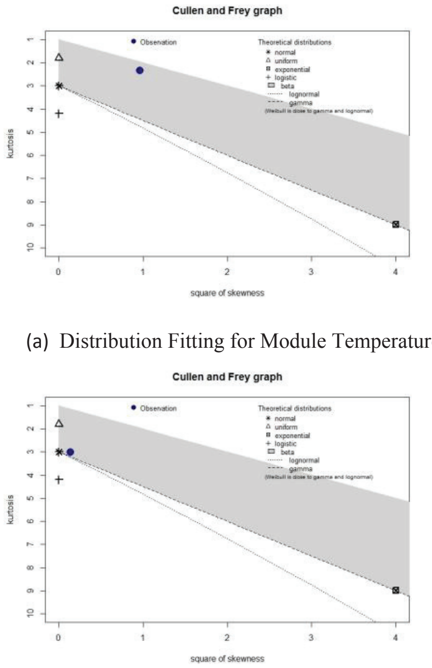


Fig. 7. Fitting Ambient Temperature using MLE.



(a) Distribution Fitting for Module Temperature

(b) Distribution Fitting for Ambient Temperature

Fig. 8. Cullen Frey Graphs for 2 datasets

E. Evaluate the Integrity of Data

The descriptive parameters for PV module temperature data is illustrated in Fig. 8(a) using a Cullen and Frey Graph that plots kurtosis against skewness. While Kurtosis is a measure of how heavy-tailed the distribution of a given data is, Skewness is a measure of its symmetry [29]. For example, the PDF of an ideal normally distributed data has a Skewness of 0 and a Kurtosis of 3. It can be seen from the figure that the PDF of the considered dataset, denoted by a blue colored filled circle, is likely to show a fit for beta distribution denoted by the shaded region. A similar Kurtosis-Skewness plot for ambient temperature is shown in Fig. 8(b), from which it can be seen that the data has a better fit for normal distribution, which is in agreement with the MLE-fitted model in Fig. 7. Conducting a similar case-by-case analysis on other datasets, it was observed that irradiance showed better fit to lognormal while site energy was closer to beta distribution. It can, hence, be concluded that even after processing, the inherent properties of the data are not modified, keeping its integrity intact.

IV. Conclusion

This paper is one of the first attempts at proposing a systematic roadmap for locally processing and preparing the multivariate time-series data of distribution grid-tied PV sites by leveraging lightweight analytics such as descriptive and diagnostic methods. Further, this approach aligns with the emerging distributed computing paradigms like Edge and Fog. The roadmap works on raw, collected data, checks it for format, structure and quality, and also explores its properties that could be later used during higher-level analytics. This roadmap is catered specifically to site-level analytics that include but are not limited to performance monitoring, component-level reliability analysis, voltage profile analysis at PCC, and PV generation forecasting modeling. The roadmap comprises five key steps: identify the objective, collect the required raw data, estimate the missing data, explore the data for its structure and properties using correlation and MLE, and finally compare the processed data's distribution to that of the raw data to ensure its integrity is not compromised. As a future work, the proposed roadmap will be applied to other site-level analytics and then integrated with more computationally intensive applications based on cloud or Hadoop. Further, the data processed using roadmap will be used to calculate site performance metrics and the results will be discussed.

References

[1] A. Sanjab, W. Saad, I. Guvenc, A. Sarwat, and S. Biswas, "Smart grid security: Threats, challenges, and solutions," in *arXiv:1606.06992*, Jun. 2016.

[2] I. Parvez, M. Jamei, A. Sundararajan, and A. I. Sarwat, "Rss based loop-free compass routing protocol for data communication in advanced metering infrastructure (ami) of smart grid," in *2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, Dec 2014, pp. 1–6.

[3] A. Anzalchi and A. Sarwat, "A survey on security assessment of metering infrastructure in smart grid systems," in *IEEE Southeast Conference*, Fort Lauderdale, 2015.

[4] OpenFog, "Openfog reference architecture for fog computing," *OpenFog Consortium Architecture Working Group*, 2017.

[5] L. Wei, A. H. Moghadasi, A. Sundararajan, and A. Sarwat, "Defending mechanisms for protecting power systems against intelligent attacks," in *Proc. IEEE 10th SoSE Conf.*, San Antonio, the United States, May 2015.

[6] M. Jamei, A. Sarwat, S. Iyengar, and F. Kaleem, "Security breach possibility with rss-based localization of smart meters incorporating maximum likelihood estimator," in *International Conference on Systems Engineering*, Las Vegas, 2015.

[7] I. Parvez, A. Sundararajan, and A. Sarwat, "Frequency band for han and nan communication in smart grid," in *IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, Orlando, Dec. 2014.

[8] SunSpec, "Best practices in solar performance monitoring," *SUNSPEC Alliance Technical Report*, 2014.

[9] R. Shyam, H. B. B. Ganesh, and S. S. K. et al, "Apache spark a big data analytics platform for smart grid," *Smart Grid Technologies, Procedia Technology*, 2015.

[10] S. Joseph, E. A. Jasmin, and S. Chandran, "Stream computing: Opportunities and challenges in smart grid," *Smart Grid Technologies, Procedia Technology*, 2015.

[11] V. Silva, F. Rodriguez, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions in Power Systems*, vol. 4, no. 13, pp. 596–602, Jun. 2005.

[12] CSCC, "Deploying big data analytics applications to the cloud: Roadmap for success," *Cloud Standards Customer Council (CSCC) Report*, May 2014.

[13] A. Beres, B. Genge, and I. Kiss, "A brief survey on smart grid data analysis in the cloud," in *8th International Conference on Interdisciplinarity in Engineering (INTER-ENG)*, Oct. 2014.

[14] H. Daki, A. E. Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big data management in smart grid: concepts, requirements and implementation," *Journal of Big Data*, vol. 4, no. 13, pp. 1–19, Apr. 2017.

[15] C. S. Lai and L. L. Lai, "Application of big data in smart grid," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2015.

[16] SWECO, "Smart grid and big data analytics," *SWECO Report*, 2016.

[17] M. S. de Cardonaa and L. M. Lopezb, "Performance analysis of a gridconnected photovoltaic system," *Pergamon Energy*, 1998.

[18] PVPS, "Analytical monitoring of grid-connected," *PhotoVoltaic Power Systems Programme Report*, 2014.

[19] M. Bayrakcia, Y. Choib, and J. R. S. Brownsona, "Temperature dependent power modeling of photovoltaics," in *ISES Solar World Congress*, Oct. 2014.

[20] A. M. Omar, M. Z. Hussin, S. Shaari, and K. Sopian, "Energy yield calculation of the grid connected photovoltaic power system," *Computer Applications in Environmental Sciences and Renewable Energy*, 2014.

[21] S. V. Buuren, J. P. L. Brand, K. Groothuis-Oudshoorn, and D. B. Rubin, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation*, 2006.

[22] S. V. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, pp. 1–68, 2007.

[23] A. A. Beaujean, "R package for baylor university educational psychology quantitative courses," *R CRAN*, Aug. 2012.

[24] M. L. Delignette-Muller, R. Puillot, and J. B. Denis, "Fitting parametric distributions using r: the fitdistrplus package," *useR! Conference Report*, 2009.

[25] L. Wei, A. Sundararajan, A. Sarwat, S. Biswas, and E. Ibrahim, "A distributed intelligent framework for electricity theft detection using benford's law and stackelberg game," in *Resilience Week*, 2017.

[26] Greene, "Maximum likelihood estimation," Nov. 2010.

[27] P. E. C. of Science, "Maximum likelihood estimation," *Probability Theory and Mathematical Statistics*.

[28] E. Zivot, "Introduction to maximum likelihood estimation," *Technical Presentation*.

[29] A. Cullen and H. Frey, *The Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York, NY: Plenum Press, Plenum Publishing Corporation, 1999.