A Data-Driven Approach for Detection and Estimation of Residential PV Installations

Xiaochen Zhang, *Student Member, IEEE*, and Santiago Grijalva, *Senior Member, IEEE*

***Abstract*—The number of photovoltaic (PV) systems in the electric grid is growing at an unprecedented speed. This is rapidly transforming the ways in which the traditional distribution grid is being planned and operated. A problem faced by utilities is that, in many cases, the PV system installed does not correspond to the size or type filed with the installation permit, or simply the installation took place without a permit. In order to maintain grid reliability and safety, utilities must be able to detect and monitor all PV installations in their network. This paper proposes a data-driven approach for the detection, verification, and estimation of residential PV system installations. We use a change-point detection algorithm to screen out abnormal energy consumption behaviors including unauthorized PV installations. Then the existence of the unauthorized PV installation is further verified through a statistical inference known as permutation test with Spearman's rank coefficient. The proposed hypothesis test takes the customer's load profiles before and after the detected change-point as inputs, which are estimated through Gaussian kernel density method. Finally, the local cloud cover index is integrated with smart meter measurements to estimate the size of the PV system. The proposed method has been tested and validated with actual smart meter measurements under several scenarios.**

***Index Terms*—Change-point detection, Gaussian kernel smoothing, permutation test, parameter estimation, PV system, Spearman's rank coefficient.**

I. Introduction

A VARIETY of distributed energy resources (DER) such as photovoltaic (PV) systems, micro turbines, and electrical vehicles (EV) are being connected to the grid [1]. According to the Hawaiian Electric Company (HECO), in 2015 one in eight of HECO's 450,000 customers has a residential PV system. As the speed of residential PV adoption continues to accelerate, in a high PV penetrative environment, utilities are facing technical problems related to overvoltage, frequency control, and back feeding flow [2], as well as issues such as a rapid decrease in revenue. In order to manage these new challenges, it is critical for utilities to gain visibility of all plugged-in PV systems, especially at the residential level.

Unauthorized PV installations may create safety risks, and lack of visibility may result in incorrect planning and operation, including over-voltages, back-feeding, and in the worst case scenario, damaging system equipment such as transformers, voltage regulators, as well as customer's appliances [3], [4]. In order to facilitate the adoption of residential PV systems and to minimize risks, utilities enforce regulations and permits for residential PV systems. In California, Hawaii, and other states, it is required by law [5], [6] that customers should obtain necessary permits from a permit agency before any PV system installation. According to the DOE's report on smart grids 2014 [7], massive adoption of PV systems will lower the utilities' revenue, which in return increases the bill for non-solar customers. In Arizona, a fixed charge for new customers who sign a contract with a solar energy provider was recently implemented [8], which leads to similar debate about solar interconnection fees in many states.

There are various reasons for unauthorized or incorrectly registered PV systems: a) Owner decided not to apply for a permit to avoid permit fees [9], b) Regulations were required after the system was installed, c) Lack of awareness by the owner of diverse permitting rules by country, state, city and often zonal regulations, d) Different rules depending on the size and type of PV installation can make the owners believe they do not need a permit, e) Changes in property ownership including transfers, f) Multiple systems installed or future additions at the same premises, g) Incorrect third party handling of the permit application, and h) Data entry and data maintenance errors. In 2014 Hawaii, the system with the highest penetration of PV in the U.S., recognized a large number of unauthorized PV installations [3] and prompted a specific program trying to reduce the number of these systems. In North Belgium, the number of unauthorized PV systems has exceeded that of the PV system installed under the local certificate due to the introduction of the grid fee in 2013 [9]. This creates a serious problem for the operation and long-term planning of distribution systems.

An effective and efficient PV system detection and estimation algorithm can be proved to be of significant services to utilities for safety, reliability, and revenue reasons [10]. If not accurately modeled and managed, the fast adoption of PV in the distribution system can put the system security and reliability at risk. Traditional distribution networks are designed for one-directional power flow. High penetration of PV can lead to reverse power flow along the distribution feeders [11] and cause system protection failures. In addition, PV output is heavily influenced by sky cloud cover

Manuscript received September 1, 2015; revised January 28, 2016; accepted March 8, 2016. Date of publication April 21, 2016; date of current version August 19, 2016. Paper no. TSG-01026-2015.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: x.zhang@gatech.edu; sgrijalva@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2016.2555906

and can be highly variable, resulting in numerous energy spikes, transient over-voltage [4] and increased transformer tap-change operations.

Many researchers have studied the impacts and risks of PV on distribution systems [12]–[17]. However, the detection and monitoring of residential PV systems has not been the focus of the studies and related research. This paper proposes a data-driven approach to detect and monitor unauthorized and misfiled PV systems by implementing advanced data mining algorithms on smart meter data stream.

Thanks to the significant investment on smart meters and their related advanced metering infrastructures (AMI) in the past few years, the database populated with smart meter measurements is starting to play a very important role in utilities' daily operations, such as enhanced load forecast [18], load modeling, demand response, and load profiling [19], [20]. In this paper, we show that the historical data collected by smart meters can also help utilities detect unauthorized or misfiled PV systems in order to enhance their customer models. Better models and accurate databases result in significant operational benefits to the utility. The proposed method consists of three steps:

- Step 1: Unauthorized PV system screening
- Step 2: PV system verification test
- Step 3: PV size estimation.

Due to computational efficiency, on the first step, we detect energy consumption abnormalities among all customers using a recently developed change-point detection algorithm [21], which returns the abnormalities as change-points in the energy consumption time-series data. On the second step, we estimate the typical load profiles (TLP) before and after the change-point using Gaussian kernel density estimation, which filters out noises that result from the customer's random behaviors. We construct a statistical inference using the permutation test with Spearman's rank coefficient to verify whether the change-point is caused by an unauthorized PV installation. Once an unauthorized PV installation has been confirmed by the statistical inference, we further estimate the size (rated power) of the detected PV system using the local cloud cover index (CCI). CCI is a numerical measure of the fraction of the sky obscured by clouds [22]. The proposed method has been validated on realistic system data sets, where all load components including PV outputs are recorded through separate meters.

The rest of the paper is structured as follows: Section II formulates the mathematic model of the smart meter data as well as the structure of the proposed 3-step method. Section III elaborates the adopted change-point detection algorithm based on relative density-ratio estimation. Section IV shows how to form the Gaussian-kernel-based TLP and construct statistical inference to verify the existence of an unauthorized PV system. Section V discusses the strong correlation between the PV output and local CCI and how to incorporate CCI for PV size estimation. Section VI shows a real case study of the proposed method under three different scenarios. Section VII concludes the paper and discusses future research opportunities on PV system detection and estimation.

II. Problem Formulation

In this section, we discuss the organization of the smart meter measurements used in this study and formulate the residential PV detection problem as a combination of a change-point detection problem, a statistical inference, and a parameter estimation problem.

A. Smart Meter Time Series Data

The data set used in this study corresponds to a set of 15-minutes-resolution smart meter readings from hundreds of homeowners from a U.S. city, in 2013. Around 40 of these homeowners have home solar systems installed, and the corresponding 15-minutes-resolution PV outputs for each house are recorded through separate meters. The energy consumption and PV output data from these 40 PV-equipped houses are used in our study.

We model the smart meter historical readings as time series streamed data, with the frequency of f readings per day ($f = 96$ in this study). Let $y(t_{d,i})$ denote the ith reading for day d, and let

$$\boldsymbol{D}(d) \coloneqq \left[\, y\left(t_{d,1}\right)\, y\left(t_{d,2}\right) \cdots y\left(t_{d,f}\right) \,\right]^T \in \mathbb{R}^f \tag{1}$$

denote the sequence of smart meter readings for day d. We batch the daily measurements into a data bundle $\boldsymbol{Y}(d)$ as in equation (2), where the time window is k days. Then,

$$\boldsymbol{Y}(d) \coloneqq [\, \boldsymbol{D}(d)\, \boldsymbol{D}(d{+}1) \cdots \boldsymbol{D}(d{+}k{-}1)\,] \in \mathbb{R}^{f\times k} \tag{2}$$

corresponds to all the smart meter readings starting from day d to day $(d{+}k{-}1)$. The data bundle $\boldsymbol{Y}(d)$ is later used as input for the change-point detection algorithm. This is illustrated in Fig. 1, where the change-point detection algorithm compares the differences between every two adjacent data bundles.

B. PV Detection Problem Formulation

The residential PV system installation detection can be formulated as a change-point detection problem. Let us consider a PV system installed at day $(d{+}k)$. The PV energy output will be reflected on the smart meter measurements of the customer. As a result, the smart meter readings or the data bundles before and after the PV installation date (e.g., $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d{+}k)$) must be dissimilar. We use Pearson divergence (PE divergence) to measure the dissimilarity between two different data bundles $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d\,+\,k)$, see equation (3) [23]. The change-point is detected based on the PE divergence score tested on every adjacent pair of data bundles, as shown in Fig. 1. Let us assume P and P' to be the distribution of the data in data bundles $\boldsymbol{Y}(d)$ and $\boldsymbol{Y}(d{+}k)$, then $PE(P||P')$ is the PE divergence between distribution P and P', which can be computed using (3).

$$PE(P||P^{'}) \coloneqq \frac{1}{2}\int p'(\boldsymbol{Y})\left(\frac{p(Y)}{p'(Y)}-1\right)^2 d\boldsymbol{Y}, \tag{3}$$

where $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are the probability density functions of the two distributions P and P'.

In order to verify the existence of an unauthorized PV system, we construct a hypothesis test based on the energy output

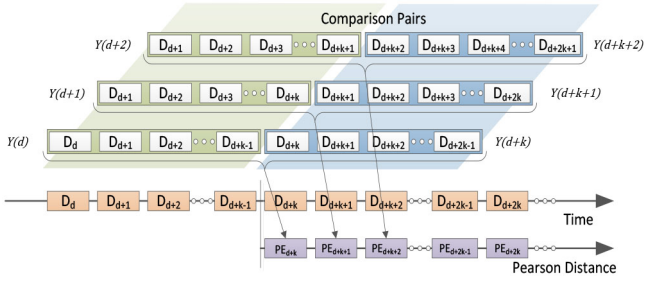

Fig. 1. Time series data formulation.

behaviors of a PV system. Finally, we determine the size of the detected PV system by solving a parameter estimation problem using the local weather information.

III. PV Change-Point Detection

A. Change-Point Detection

Change-point detection or change-point analysis is a powerful tool used to detect abrupt changes in time series data. This method has been widely applied in many areas such as climate change [24], image processing [25] and financial economics [26]. Most change-point detection methods can be categorized into two classes: real-time detection and off-line detection. In this paper, we adopt a recently developed off-line detection method that uses relative density-ratio estimation to detect abnormalities in customer energy consumption [21]. For a time series data set, the change-point detection algorithm can detect various changes, such as jumping mean, scaling variance, switching covariance, or even varying frequency caused by PV installation.

B. Relative Density-Ratio Estimation

The change-point detection algorithm developed in [21] is used in this paper due to its efficiency and non-parametric nature. In equation (3), since the true $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are unknown, the estimated densities $\hat{p}(\boldsymbol{Y})$ and $\hat{p}'(\boldsymbol{Y})$ are used to calculate the PE divergence. In the relative density-ratio estimation method, instead of estimating two distributions $\hat{p}(\boldsymbol{Y})$ and $\hat{p}'(\boldsymbol{Y})$ respectively (a harder problem), we only estimate one statistic, the density-ratio $g(\boldsymbol{Y};\boldsymbol{\theta}) = \hat{p}(\boldsymbol{Y})\ /\hat{p}'(\boldsymbol{Y})$, through Gaussian kernel model [27]

$$g(\boldsymbol{Y};\boldsymbol{\theta}) = \sum_{l=1}^{n} \theta_l K(\boldsymbol{Y}, \boldsymbol{Y}_l), \tag{4}$$

where $\boldsymbol{\theta}$ is an n dimensional parameter to be learnt from the data samples so that the PE divergence between $p(\boldsymbol{Y})$ and $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ is minimized; and $K(\boldsymbol{Y}, \boldsymbol{Y}_l)$ is the Gaussian kernel function evaluated at $\boldsymbol{Y}_l$.

After the density-ratio estimator $\hat{g}(\boldsymbol{Y})$ is computed using the estimated $\widehat{\boldsymbol{\theta}}$, the PE divergence can be constructed as equation (5) [15].

$$\widehat{PE} = -\frac{1}{2n}\sum_{j=1}^{n}\widehat{g}\left(\boldsymbol{Y}'_j\right)^2 + \frac{1}{n}\sum_{j=1}^{n}\widehat{g}\left(\boldsymbol{Y}'_j\right) - \frac{1}{2}. \tag{5}$$

If we consider the α-relative PE-divergence PE_α for $0 = \alpha < 1$, the symmetrized PE divergence is given as

$$PE_\alpha\left(P||P^{'}\right) + PE_\alpha\left(P'||P\right) \tag{6}$$

where $PE_\alpha(P||P') = PE(P||\alpha P + (1-\alpha)P')$ and α is called the "smoother" as α gets larger [28].

According to Reference [21] the introduction of a relative density-ratio provides a solution for the unbounded density-ratio for better estimation. The adopted density-ratio estimation method is also known as relative unconstrained least-squares importance fitting (RuLSIF). Compared with other change-point detection methods, RuLSIF has several advantages for PV installation detection. First, RuLSIF is parameter-free. We only need to control the time window length k, as shown in equation (2). Second, RuLSIF estimates one density-ratio instead of two density functions, which is computational efficient and substantially easier [21]. Finally, RuLSIF is known for its optimal non-convergence rate and robustness compared with other time-series-based methods [21].

IV. PV Detection Identification

The change-point detection algorithm discussed in Section III can detect abnormalities in customer's energy consumption history caused by the PV installation. However, other customer behaviors such as introducing a new EV or a sudden drop of temperature will also cause abrupt energy consumption abnormalities and thus be detected and marked with a change-point. As a result, once an abnormality is detected, a statistical inference must be constructed to further verify whether the sudden change of customer behavior is caused by the installation of a PV system.

A. Typical Load Profile

The typical load profile, which summarizes the customer's energy consumption pattern, plays a fundamental role in utility's daily operation. In this paper, we introduce a daily TLP to compare a customer's power consumption patterns before and after the change-point. Let us assume the smart meter collects f readings per day. The daily TLP of a specific customer can be represented by a vector $\boldsymbol{V}_{TPL} \in \mathbb{R}^f$. Given a time window of n days, the TLP for the customer can be computed using equation (7).

$$\boldsymbol{V}_{TLP}(\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}(i) \in \mathbb{R}^f. \tag{7}$$

Due to the fact that most smart meters are installed at the residential level, the random behaviors of homeowners may cause spikes along their energy consumption history. These spikes introduce significant noises to TLP estimation in equation (7). In order to filter out unnecessary noises, we use the Gaussian kernel density method to estimate the TLP. Kernel density estimation is a non-parametric algorithm originally used for probability density function estimation. Since kernel density estimators asymptotically converge to any density

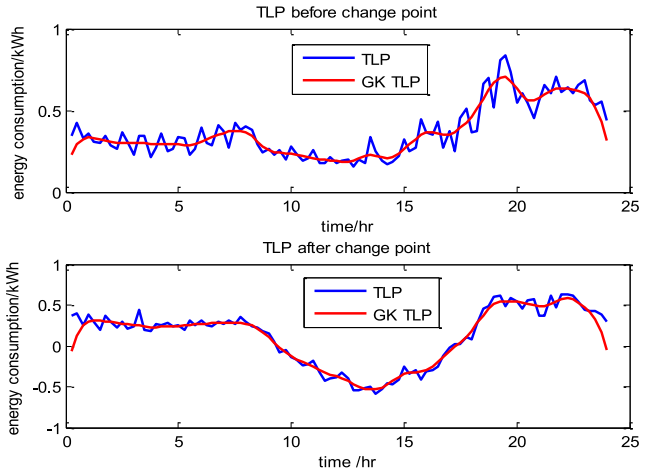

Fig. 2. Gaussian kernel based TLP.

function with sufficient samples, it is a very general estimation method [29] and is robust for a variety of TLP shapes. Compared with simply taking the mean value in (7), Gaussian kernel approach returns a much smoother TLP with less noise and requires less space to store. In our study, the TLP curve is treated as a probability density function and Gaussian kernels are used to estimate the TLP. The estimated density function $\hat{f}(x)$ with m kernels can be computed by equation (8):

$$\widehat{f}(x) = \sum_{i=1}^{m} w_i K(x - x_i), \tag{8}$$

where $K(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-x_i)^2}{2\sigma^2})$ is the Gaussian probability density function with mean x_i and variance σ^2, w_i is the weight of each Gaussian kernel that satisfies $\sum w_i = 1$.

Fig. 2 shows two TLPs of a customer before and after a PV system was installed. The blue curves are TLPs computed using equation (7). The red curves correspond to TLPs smoothed by the Gaussian density estimation method. It is clear that the Gaussian-shaped kernel serves to smooth out the noises in TLP.

B. Statistical Hypothesis Test With Spearman's Rank

When an abnormal customer behavior is detected, it is crucial for utilities to verify whether the abnormality is caused by a PV installation. Instead of issuing a field work order and on-site inspection, we construct a TLP-based hypothesis test to verify the existence of an unauthorized PV system. Specifically, we construct the null hypothesis (H_0) as the following statement: "There is no unauthorized PV system installed by the customer." In other words, we generally assume that there is no unauthorized PV installation unless evidence strongly indicates otherwise.

Similar to a customer's TLP, $\boldsymbol{V}_{TPL} \in \mathbb{R}^f$, we define $\boldsymbol{V}_{PV} \in \mathbb{R}^f$ as a standard TLP of a local PV system. $\boldsymbol{V}_{PV}$ records the standard daily energy output of the local PV systems with rated power equals to 1 kW. Let $\Delta\boldsymbol{V}_{TLP} \in \mathbb{R}^f$ denote the difference of TLPs before and after the change-point. If the detected change-point is caused by an unauthorized PV system, we have $\Delta\boldsymbol{V}_{TLP} = p\boldsymbol{V}_{PV}$, where p is the size of the unauthorized PV system. Otherwise, we will not be able to find a constant p so that $\Delta\boldsymbol{V}_{TLP} = p\boldsymbol{V}_{PV}$ is true.

Let us define

$$\Delta\boldsymbol{V}_{TLP} = \boldsymbol{X} = (x_1,\ x_2, x_3, \ldots, x_f) \tag{9}$$

$$\boldsymbol{V}_{PV} = \boldsymbol{Y} = \left(y_1,\ y_2, y_3, \ldots, y_f\right) \tag{10}$$

Then, the original hypothesis test can be rephrased as:

$$\begin{cases} H_0\text{: } \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are not positively correlated} \\ H_1\text{: } \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are positively correlated} \end{cases} \tag{11}$$

C. Spearman's Rank and Permutation Test

Pearson product-moment correlation (Pearson's r) and Spearman's rank correlation coefficient (Spearman's rank) are the most commonly used metrics to quantify the correlation between two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ [30]. However, the difference between the two methods lies in that Pearson's r assumes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are normally distributed, while Spearman's rank does not have any requirement on the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this paper, we adopt Spearman's rank (r_s) due to the fact that the distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in (9) and (10) are not normal. The Spearman's rank coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be computed using equation (12) [31].

$$r_s = 1 - \frac{6(\sum d_i^2)}{n(n^2-1)} \tag{12}$$

where r_s is the Spearman's rank coefficient ($-1 = r_s = 1$). When $|r_s|$ is close to 1, it indicates a strong linear relationship between the two distributions, and 0 otherwise. n is the number of $(x_i,\ y_i)$ pairs in observation which, in our case, $n = f$, and $d_i = x_i - y_i$. Since r_s quantifies the strength of the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, an interpreting table developed by Hinkle [31] is usually used for interpreting the physical meaning of r_s [32].

In our hypothesis test, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not normally distributed, we cannot use a t-test to acquire an accurate p-value through the student distribution. Instead of using t-test, we adopt the permutation test. Permutation test (a.k.a. randomization test) is a very general approach to test a statistical hypothesis, where the distribution of the observations under the null hypothesis need not be known to obtain the p-value [33].

The existence of an unauthorized PV system will drive r_s close to 1. Hence, we can further rephrase the original null hypothesis in (11) as H_0:$r_s = 0$. Next, we select a significance level α and compute the p-value through the permutation test. For f pairs of (x_i, y_i) listed in (9-10), the total number of permutation sets is 2^f. Let $r_{s,i}$ stand for the Spearman's rank coefficient for permutation set π_i and $r_{s,0}$ for the observed Spearman's rank coefficient of permutation π_i. Since we want to test whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are positively correlated, the permutation test is no longer a two-tailed test but an upper tailed test. Therefore, the corresponding test procedure can be decomposed using the following three steps:

Step 1 Generate all possible permutation sets [34]:

$$\pi_1,\ \pi_2, \ldots, \pi_{2^f}.$$

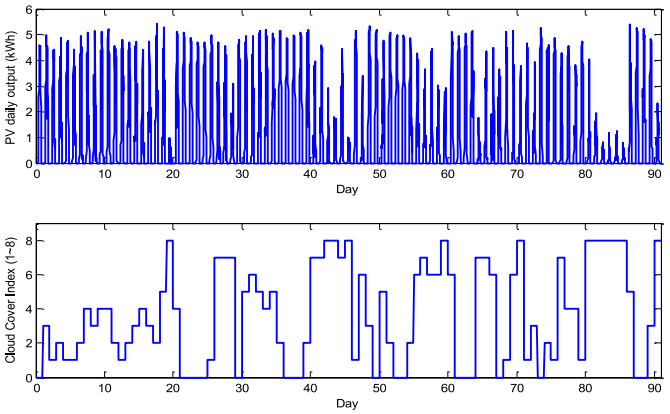

Fig. 3. PV output and corresponding CCIs for 91 days.

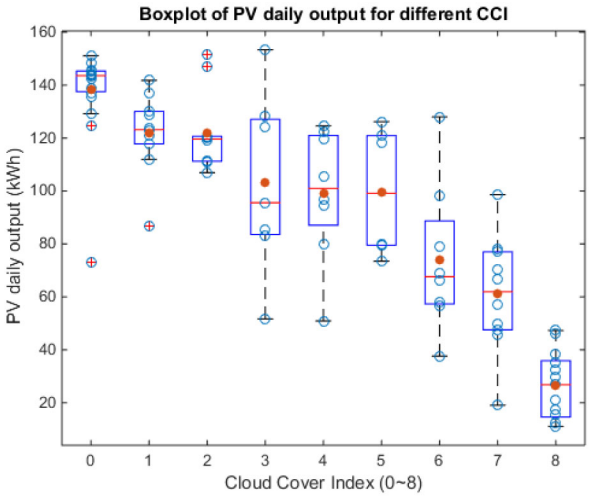

Fig. 4. Boxplot of PV daily output vs. local CCI.

Step 2 Compute the Spearman's rank for all sets:

$$r_{s,1}\ r_{s,2}, \ldots,\ r_{s,2^f}.$$

Step 3 Construct an empirical cumulative distribution [33]:

$$\widehat{p}\left(r_s = r_{s,0}\right) = \frac{1}{2^n}\sum_{i=1}^{2^f} 1\left(r_{s,i} = r_{s,0}\right) \tag{13}$$

where $\hat{p}$ is the cumulative density function of the estimated Spearman's rank coefficient. $1(s)$ is an indicator function which takes value 1 if statement s is true and 0 otherwise. In practice, when the number of $(x_i,\ y_i)$ pairs is generally large (in our case $f = 96$), it is difficult to generate all possible 2^f permutations. As a result, bootstrap sampling must be implemented. For the significance level of $\alpha = 0.05$, according to Reference [33], 10,000 bootstrap samples are recommended.

Given a preset significance level α, we reject the null hypothesis, if $\hat{p} = \alpha$. In other words, $\hat{p} = \alpha$ indicates that there is a very good chance the detected customer has an unauthorized PV system installed.

V. PV Behavior Analysis and Size Estimation

Among all PV parameters, the size or the rated power of the PV system p, is the most important. However, as a parameter estimation problem, a good estimation of p is difficult when only smart meter measurements are available. This is because the PV output is strongly affected by the weather condition such as local solar irradiance and cloud cover. CCI obtained from satellite images contains information on cloud amount and their optical thickness [35]. To be more specific, CCI is defined as an integer ranges from 0 to 8, where 0 stands for clear-sky day and 8 stands for heavily clouded day.

In this section, we select a residential PV system and record its output for 91 consecutive days, as shown in Fig. 3-1. The local CCIs for the corresponding days are shown in Fig. 3-2. We can see that on high CCI days, the PV output is generally small, and vice versa. The correlation between the PV daily output and the CCI is -0.8554, which indicates high linear correlation between the two. This meets our expectation that more cloud in the sky leads to lower PV output.

In order to visualize the correlation between CCI and PV output, the boxplot (a.k.a. box and whisker diagram) of PV daily output condition on the CCIs is shown in Fig. 4 using the previous data. According to the boxplot definition [36], the central red mark is the median, the edges of the box are the first and third quartiles, and the red-cross stands for outliers. From Fig. 4, we can see that the PV output variance increases as the CCI increases from 0 and decreases when CCI approaches to 8. This phenomenon can be explained by the fact that when CCI is in the middle range, the sky is partially covered by cloud, and the passing of clouds above the solar panel may lead to huge variance on PV output. In order to obtain an accurate PV size estimation, only days with low CCI can be used, where the PV output has small variance and is near to its rated output.

Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ stand for the smart meter readings before and after the PV installation respectively. Let $\widetilde{\boldsymbol{D}}_2$ be an adjusted $\boldsymbol{D}_2$ according to local CCIs and radiance. For a specific day k, $\widetilde{\boldsymbol{D}}_2(k)$ can be computed using (14).

$$\widetilde{\boldsymbol{D}}_2(k) = \boldsymbol{D}_2(k) - p \times p_{CCI}(k) \times \boldsymbol{V}_{PV}, \tag{14}$$

where p is the size of the PV system, $p_{CCI}(k)$ is the adjustment coefficients related to the local CCI and radiance on day k, which increases as CCI increases. In practice, p_{CCI} can be estimated based on empirical distribution of PV output condition on the local CCI. Then, the PV size estimation problem becomes choosing the best constant p that minimizes (15).

$$min:\ \left\| \boldsymbol{V}_{TLP}(\boldsymbol{D}_1) - \boldsymbol{V}_{TLP}(\widetilde{\boldsymbol{D}}_2) \right\|^2 \tag{15}$$

where $\boldsymbol{V}_{TLP}(\boldsymbol{D}_1)$ and $\boldsymbol{V}_{TLP}(\widetilde{\boldsymbol{D}}_2)$ stand for the typical load profiles computed using $\boldsymbol{D}_1$ and $\widetilde{\boldsymbol{D}}_2$.

VI. Real Case Analysis Results

In this section, we investigate the performance of our method on real data sets. The data contain a rich source of disaggregated customer energy consumption. In order to show the robustness of the proposed method, a representative subset

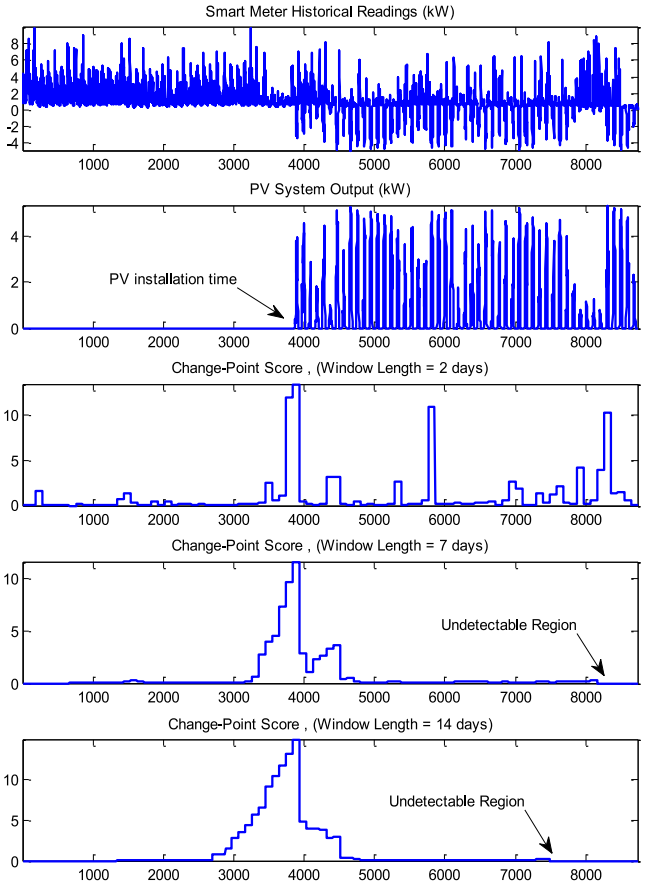

Fig. 5. Change-point detection screening for an unauthorized PV installation.

of the data described in Section II is used which includes three distinct scenarios:

- Scenario 1: Customer A has installed an unauthorized PV system;
- Scenario 2: Customer B has bought a new EV and experienced a major weather change;
- Scenario 3: Customer C has no abnormal behavior.

We expect that our proposed algorithm will only identify the customer A in scenario 1, where an unauthorized PV system exists.

A. Change-Point Detection Screening

The proposed change-point detection algorithm will pick up energy abnormality efficiently when historical smart meter data are available. The real case study shows that only customer C in scenario 3, who does not have any abnormal energy consumption behaviors, can pass our change-point detection screening.

- Scenario 1: An unauthorized PV system is installed

In scenario 1, a smart meter monitored the aggregated power consumption of customer A for 91 consecutive days with 8736 measurements, as shown in Fig. 5-1. The negative values in Fig. 5-1 stand for the PV system back feeding to the grid. The unauthorized PV system was installed on the 41th day and the PV output is recorded by a separate meter as shown in Fig. 5-2. The reading of this meter is invisible to the local utility.

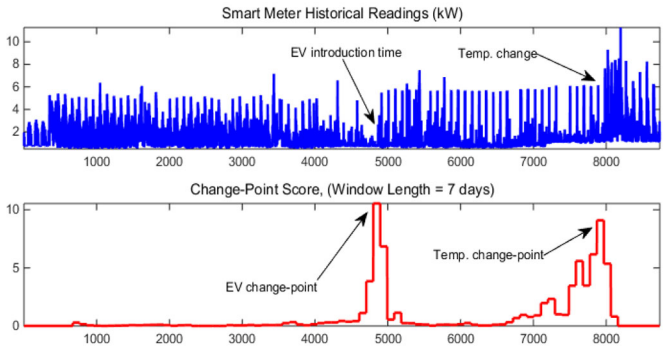

Fig. 6. Change-point detection screening for a new EV and temparture changes.

Given the parameter-free nature of RuLSIF, analysts only need to determine the estimation window length k as in equation (2). The performance of the change-point detection algorithm relies on a proper choice of k. The algorithm takes the aggregated data in Fig. 5-1 as inputs, and returns the PE divergence scores in Fig. 5-3, Fig. 5-4 and Fig. 5-5, each with a different time window length (2 days, 7 days and 14 days). Due to the smart meter data structure formulated in equation (1) and (2), the algorithm will leave two blind detection periods located at the beginning and the end of the time series stream as shown in Fig. 5-4 and Fig. 5-5. The undetectable period length equals to the length of the estimation window k. In other words, the algorithm cannot detect newly installed PV system until k days after the initial installation. From Fig. 5-3, Fig. 5-4, and Fig. 5-5, we see that shorter estimation window will enable the detection of some short term changes in the customer behavior and also minimize the undetectable period at the expense of lower index stability. However, the installation of a PV system is not likely to be a short-term activity, a longer estimation window can increase the robustness of the algorithm. Compared with Fig. 5-4 and Fig. 5-5, Fig. 5-3 is generated with a much shorter time window and its PE divergence score is less stable. Therefore, a balance must be maintained when choosing a proper time window. In our study, we set an appropriate estimation window length as 7 days.

- Scenario 2: A new EV and load fluctuations caused by weather changes

In Scenario 2, no unauthorized PV is presented during the 91-day study period. However, a new EV was introduced at the 51th day and the customer also experienced a sudden temperature change at the 82th day. From Fig. 6, the change-point detection algorithm picks up two change-points when the EV was introduced and when the temperature fluctuated.

- Scenario 3: Customer without abnormal behaviors

In Scenario 3, there is no PV, EV introduction or huge temperature fluctuations, as shown in Fig. 7. The change-point detection algorithm does not pick up any significant change-point and the PE divergence scores are consistently below 2.5. As a result, the customer in scenario 3 passes our change-point screening test (no abnormality has been detected).

B. PV System Verification

On the previous step, the only customer in scenario 3 passes the change-point screening test, which leaves us with

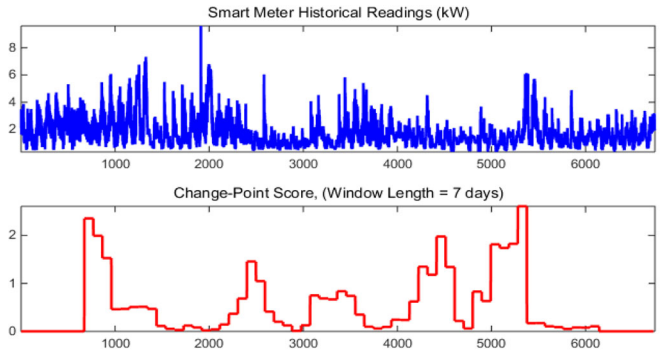

Fig. 7. Change-point detection screening for customer without abnormal behaviors.

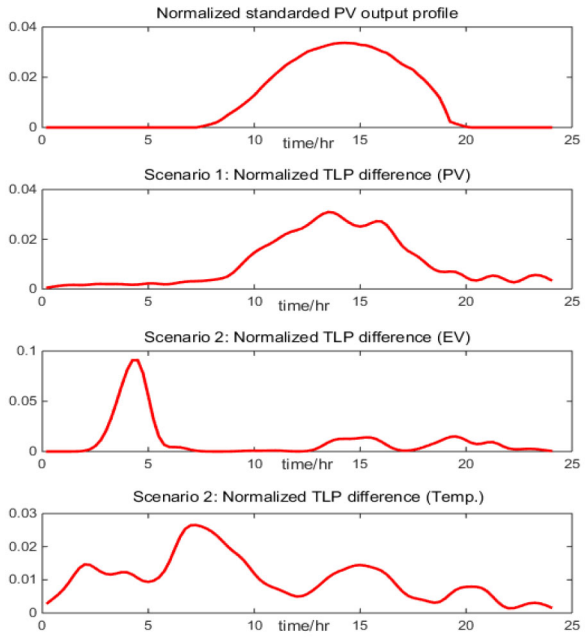

Fig. 8. Gaussian kernel based typical load profiles.

customers A and B of scenarios 1 and 2. On the second step, we use the statistical inference constructed in Section IV to identify customers without unauthorized PV system, but fail the screening test as in scenario 2. We first create the Gaussian kernel based TLPs before and after each detected change-point and compute their differences. Next, we conduct the permutation test with Spearman's rank coefficient to verify the existence of an unauthorized PV system.

In this study, the standard local PV system output profile $\boldsymbol{V}_{PV}$ is approximately estimated by taking the normalized output of 40 local PV systems on a cloud free day, as shown in Fig. 8-1. The $\Delta \boldsymbol{V}_{TLP}$ for three change-points in scenario 1 and scenario 2, as shown in Fig. 8-2, Fig. 8-3 and Fig. 8-4, are computed using equation (7) and (8). In this study, we choose 10 days as the time window to create the TLPs. All the TLPs shown in Fig. 8 are normalized and smoothed by Gaussian kernel method.

Based on Fig. 8, we perform correlation strength analysis between the standard PV output $\boldsymbol{V}_{PV}$ (Fig. 8-1) and $\Delta \boldsymbol{V}_{TLP}$ of each detected change-point (Fig. 8-2, Fig. 8-3, and Fig. 8-4).

TABLE I
CORRELATION STRENTH ANALYSIS

Change Point	Pearson's r	Spearman's rank coefficient	
	r	r_s	*p-value*
Scenario 1 (PV)	0.9205	0.8351	3.9414e-26
Scenario 2 (EV)	0.1290	-0.0315	0.7609
Scenario 2 (temp.)	0.0817	-0.0754	0.4651

TABLE II
SENSITIVITY ANALYSIS

Scaling Factor r	Change-Point Detection		Permutation Test $\alpha = 0.05$	
	Detection	*Goodness*	r_s	*p-value*
100%	yes	0.3114	0.8351	3.9414e-26
90%	yes	0.3186	0.8268	3.1698e-25
80%	yes	0.3334	0.8169	3.4432e-24
70%	yes	0.3540	0.7985	1.9597e-22
60%	yes	0.3833	0.7686	6.1399e-20
50%	yes	0.4135	0.7197	1.4309e-16
40%	yes	0.4593	0.6297	6.2748e-12
30%	yes	0.5568	0.4779	8.4663e-07
20%	yes	0.8000	0.2481	0.0148
10%	no	1.5076	-0.0358	0.7291

TABLE I. lists the Pearson's r and the Spearman's rank coefficient for each change-point. In TABLE I, only the customer in scenario 1 returns high Pearson's r and Spearman's rank coefficient which strongly indicates the existence of an unauthorized PV system. Moreover, with a choice of significance level $\alpha = 0.05$, we only reject the null hypothesis in scenario 1 where the p-value is much less than α. In scenario 2, both the p-values for the EV case and temperature case are much greater than α, which means we cannot reject our null hypothesis: there is no unauthorized PV system installation for customer B in scenario 2. After the verification step, we only accept the alternative hypothesis in scenario 1, where an unauthorized PV system truly exists.

C. Algorithm Sensitivity Analysis

In order to test the robustness of the proposed algorithm, we perform a sensitivity study for both the change-point detection algorithm and the statistical inference against the PV system size. To achieve this, we need to block all other factors which may influence our result except the PV system size. As a result, we pick the same customer with the fixed energy consumption but manually scale the output of the PV system from 100% to 10% of its original output. Let C be the energy consumption of a house and S be the PV output from the home solar system. $V = C - rS$ is the energy measurement visible to us, where r is the PV size scaling factor ranges from 100% to 10%, as shown in TABLE II. The goodness of the change-point detection in TABLE II is a measurement used to quantify how confident we are about the detection [21]. The smaller the goodness value, the more reliable the detection result. No change-point is detected if the goodness of the detection is above 1. If we consider a significance level $\alpha = 0.05$, both the detection algorithm and the statistical inference show great

sensitivity. Both of them fail only in the case where we maintain the energy consumption of the customer and scale down the PV system to 10% of its original size.

D. PV Size Estimation

The third step is estimating the unauthorized PV system size. In Section V, we show that the PV output is strongly correlated with the local CCI. For simplicity, we only choose the PV output data when the local CCI is zero (clear sky days) using equation (8) and (9). For a 5kW PV system, we get the estimated PV size of 4.7912kW using the CCI information (p_{CCI} is set as 1.07 according to the empirical PV output distribution condition on local CCI and irradiance). Without CCI information, data with high CCI are also used, which leads to a PV size estimation of 2.7771kW. In fact, due to the strong correlation between PV output and CCI, it is almost impossible to get an accurate PV size estimation without the local CCI.

VII. Conclusion

In this paper, we propose a data-driven approach for residential PV detection, verification and estimation. The proposed method consists of three steps. On the first step, the unauthorized PV installation events and other abnormal customer behaviors are detected through change-point detection. On the second step, permutation test based on Spearman's rank coefficient is constructed to verify the existence of an unauthorized PV system. On the last step, the PV size is estimated with the help of the local weather information. A study using realistic data demonstrates the effectiveness and robustness of the proposed method. In the future, we would like to expand our detection and estimation to other critical load components, such as EV and temperature related loads. The disaggregation and detection of these critical load components can be proven to be beneficial for utilities to ensure safe and reliable operations.

Acknowledgment

The authors gratefully acknowledge the contributions of Dr. Y. Mei and Dr. B. Haaland of the School of Industrial Systems and Engineering at the Georgia Tech for their comments on the problem formulation.

References

[1] P. Kind, *Disruptive Challenges: Financial and Strategic Responses to a Changing Retail Electric Business*, document, Edison Elect. Inst. (EEI), Washington, DC, USA, Jan. 2013.

[2] P. Fairley. (Jan. 2015). *Hawaii's Solar Push Strains the Grid*. [online]. Available: http://www.technologyreview.com/news/534266/hawaiis-solar-push-strains-the-grid/

[3] W. Staff. (Sep. 2014). *Heco Customers Asked to Disconnect Unauthorized PV Systems*. [Online]. Available: http://khon2.com/2014/09/05/heco-customers-asked-to-disconnect-unauthorized-pv-systems/

[4] V. Schwarzer and R. Ghorbani, "Transient over-voltage mitigation: Explanation and mitigation options for inverter-based distributed generation projects," Elect.Vehicle Transp. Center, Sch. Ocean Earth Sci. Technol., Univ. Hawai'i Manoa, Honolulu, HI, USA, Tech. Rep. HNEI-02-15, Feb. 2014.

[5] *California Solar Permitting Guidebook*, Solar Permitting Work Group, Governor's Office Plan. Res., Sacramento, CA, USA, Jun. 2012.

[6] *Guide to Renewable Energy Facility Permits in the State of Hawaii*, Hawaii Clean Energy Initiative (HCEI), Honolulu, HI, USA, Apr. 2015.

[7] U.S. Department of Energy (DOE). (Aug. 2014). Smart Grid System Report. [Online]. Available: http://energy.gov/oe/downloads/2014-smart-grid-system-report-august-2014

[8] *In the Matter of Arizona Public Service Company's Application for Approval of Net Metering Cost Shift Solution*, document E-01345A-13-0248, Arizona Corp. Commission, Phoenix, AZ, USA, Dec. 2013.

[9] PHOTON. (Jul. 2013). *Flanders: Illegally Installed PV Systems Could Outnumber Systems Installed Under GC Scheme*. [Online]. Available: http://www.photon.info/photon_news_detail_en.photon?id=78701

[10] P. Denholm and R. Margolis, "Supply curves for rooftop solar PV-generated electricity for the United States," Nat. Renew. Energy Lab. (NREL), Colorado, USA, Tech. Rep. NREL/TP-6A0-44073, Nov. 2008.

[11] X. Zhang, Z. Bie, and G. Li, "Reliability assessment of distribution networks with distributed generations using Monte Carlo method," *Energy Procedia*, vol. 12, pp. 278–286, Dec. 2011.

[12] M. E. Baran, H. Hooshyar, Z. Shen, and A. Huang, "Accommodating high PV penetration on distribution feeders," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 1039–1046, Jun. 2012.

[13] M. J. Reno, K. Coogan, R. J. Broderick, J. Seuss, and S. Grijalva, "Impact of PV variability and ramping events on distribution voltage regulation equipment," in *Proc. IEEE Photovolt. Spec. Conf.*, Tampa, FL, USA, 2014.

[14] M. E. Baran, H. Hooshyar, Z. Shen, and A. Huang, "Accommodating high PV penetration on distribution feeders," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 1039–1046, Jun. 2012.

[15] P. Li, X. Yu, J. Zhang, and Z. Yin, "The H_∞ control method of grid-tied photovoltaic generation," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1670–1677, Jul. 2015.

[16] A. Samadi, L. Söder, E. Shayesteh, and R. Eriksson, "Static equivalent of distribution grids with high penetration of PV systems," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1763–1774, Jul. 2015.

[17] H. Ravindra *et al.*,"Impact of PV on distribution protection system," in *Proc. North Amer. Power Symp. (NAPS)*, Champaign, IL, USA, Sep. 2012, pp. 1–6.

[18] T. Hong, "Energy forecasting: Past, present and future," *Foresight Int. J. Appl. Forecast.*, vol. 32, pp. 43–48, Mar. 2014.

[19] X. Zhang, S. Grijalva, and M. J. Reno, "A time-variant load model based on smart meter data mining," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Minneapolis, MN, USA, 2014, pp. 1–5.

[20] X. Zhang and S. Grijalva, "An advanced data driven model for residential electric vehicle charging demand," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Denver, CO, USA, 2015, pp. 1–5.

[21] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Netw.*, vol. 43, pp. 72–83, Jul. 2013.

[22] R. E. Huschke, "Cloud cover," *Glossary of Meteorology*, 2nd ed. Boston, MA, USA: American Meteorological Soc., Aug. 2013.

[23] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Stat. Soc. Ser. B*, vol. 28, no. 1, pp. 131–142, 1966.

[24] R. Lund *et al.*, "Changepoint detection in periodic and autocorrelated time series," *J. Climate*, vol. 20, no. 20, pp. 5178–5190, Oct. 2007.

[25] P. Khandelwal, K. K. Singh, B. K. Singh, and A. Mehrotra, "Unsupervised change detection of multispectral images using wavelet fusion and kohonen clustering network," *Int. J. Eng. Technol.*, vol. 5, no. 2, p. 1401, 2013.

[26] K. J. Oh and I. Han, "An intelligent clustering forecasting system based on change-point detection and artificial neural networks: Application to financial economics," in *Proc. 34th Annu. Hawaii Int. Conf. Syst. Sci.*, Maui, HI, USA, Jan. 2001, pp. 3–6.

[27] T. Kanamori, T. Suzuki, and M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation," *Mach. Learn.*, vol. 86, no. 3, pp. 335–367, 2012.

[28] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, "Relative density-ratio estimation for robust distribution comparison," *Neural Comput.*, vol. 25, no. 5, pp. 1324–1370, May 2013.

[29] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[30] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Stat.*, vol. 42, no. 1, pp. 59–66, 1988.

[31] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioral Sciences*, 5th ed. Boston, MA, USA: Houghton Mifflin, 2003.

[32] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.

[33] J. I. Odiase and S. M. Ogbonmwan, "Correlation analysis: Exact permutation paradigm," Математички Весник, vol. 59, no. 4, pp. 161–170, 2007.
[34] Resampling (Statistics). (Dec. 2015). *Permutation Test Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests
[35] F. R. Martins, S. A. B. Silva, E. B. Pereira, and S. L. Abreu, "The influence of cloud cover index on the accuracy of solar irradiance model estimates," *Meteorol. Atmos. Phys.*, vol. 99, no. 3, pp. 169–180, 2008.
[36] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of boxplots," *Amer. Stat.*, vol. 32, no. 1, pp. 12–16, 1978.

Xiaochen Zhang received the B.S. degree from Xi'an Jiaotong University, in 2010, and dual M.S. degrees with Shanghai Jiao Tong University and Georgia Institute of Technology, in 2013. He is currently pursuing the Ph.D. degree with the Advanced Computational Electricity Systems Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, where he is also a Research Assistant. He has been involved in various research projects on power system data analytics supported by PSERC, Georgia Institute of Technology, and the Southern Company. He has also under taken a power system planning project funded by China State Grid, Inc., that integrates the Gansu Wind Farm, into the grid. His major interests lie in parallel computing for power system analysis and big-data intelligence for the future smart grid.

Santiago Grijalva received the graduate degree in electrical and computer engineering, and the M.Sc. and Ph.D. degrees from the University of Illinois at Urbana–Champaign, in 1999 and 2002, respectively. He is a Georgia Power Distinguished Professor of Electrical and Computer Engineering and the Director of the Advanced Computational Electricity Systems Laboratory with the Georgia Institute of Technology. His research interests include decentralized power system control, power system informatics and economics, and future sustainable energy systems. He is the Principal Investigator for various research projects under the Department of Energy, ARPA-E, EPRI, PSERC, and other industry and Government sponsors. From 2002 to 2009, he was a Software Architect and Consultant with PowerWorld Corporation. From 2013 to 2014, he was on an assignment to the National Renewable Energy Laboratory as the Founding Director of the Power System Engineering Center.