A Multi-Timescale Data-Driven Approach to Enhance Distribution System Observability

Yuxuan Yuan, *Student Member, IEEE*, Kaveh Dehghanpour, *Member, IEEE*, Fankun Bu, *Student Member, IEEE*, and Zhaoyu Wang, *Member, IEEE*

***Abstract*—This paper presents a novel data-driven method that determines the daily consumption patterns of customers without smart meters (SMs) to enhance the observability of distribution systems. Using the proposed method, the daily consumption of unobserved customers is extracted from their monthly billing data based on three machine learning models. In the first model, a spectral clustering algorithm is used to infer the typical daily load profiles of customers with SMs. Each typical daily load behavior represents a distinct class of customer behavior. In the second module, a multi-timescale learning model is trained to estimate the hourly consumption using monthly energy data for the customers of each class. The third stage leverages a recursive Bayesian learning method and branch current state estimation residuals to estimate the daily load profiles of unobserved customers without SMs. The proposed data-driven method has been tested and verified using real utility data.**

***Index Terms*—Observability, spectral clustering, machine learning, distribution system state estimation.**

I. Introduction

ADVANCED Metering Infrastructure (AMI) enables utilities to perform energy consumption measurement, demand-side control, tampering detection, and voltage monitoring [1]. The core element of AMI is smart meters (SMs). Compared to conventional electromechanical meters that simply record the monthly energy consumption data, SMs record the real-time load consumption of customers. Recently, a rapid growth of SMs has been observed in distribution systems. According to statistical data provided by the U.S. Energy Information Administration (EIA), the nationwide number of SMs was estimated to be 70.8 millions in 2016 with an annual growth of 6 million devices from the previous year [2]. Nonetheless, due to financial limitations and cyber-security issues, the number of SMs in many distribution networks is still limited. Hence, many utilities still rely on traditional monthly consumption data to obtain load behaviors. This lack of knowledge of real-time load behaviors inhibits effective monitoring and control of the system. One approach for solving this problem is to widely install SMs, which is cost prohibitive. As an alternative solution, we will design data-driven real-time load estimation techniques for inferring customers' behaviors [3].

In recent years, several papers have focused on load estimation, including missing data reconstruction, communication delay compensation, and unobserved customer behavior inference. The previous works in this area can be classified into two categories based on the temporal granularity of customer datasets used for model development: *Class I*: A number of articles use data with at least hourly resolution for training load estimation methods [4]–[8]. In [4], a K-means-based load estimation approach is proposed to estimate the missing measurements by using historical half-hourly energy consumption data. In [5], a truncated Fourier series representation and cluster analysis are utilized to estimate a hybrid model of consumer load during summers. In [6], several linear Gaussian load profiling techniques are employed to capture customer behaviour using SM data analysis. In [7], in addition to SM data, the context information of customers, such as operation time during the weekends and economic codes, are leveraged to allocate the respective load profiles among particular groups, utilizing a probabilistic neural network (PNN)-based approach. In [8], power flow simulation data with half-hourly temporal resolution is exploited to obtain load estimation using Artificial Neural Networks (ANN). *Class II*: Instead of using data with high temporal resolution, a number of papers estimate the hourly customer energy consumption by converting the monthly billing data into daily load profiles [9]–[11]. In [11], hourly load estimation is performed using uniform energy allocation, where the mean and variance of estimated load is adjusted in real-time utilizing supervisory control and data acquisition (SCADA) devices. In [9], typical load profiles are assigned to the unobserved customers by comparing average daily consumption values with the daily energy levels of the representative load profile obtained from observed customers. The pseudo load profiles of unobserved customers are scaled by multiplying the estimated average consumption with the corresponding load pattern. Based on the monthly energy level, the daily load profile of unobserved customer can be obtained using representative curves from statistical analysis of residential, commercial, and industrial consumers' historical data [10].

While previous works provide valuable results, many questions remain open with respect to the real-time load estimation in distribution systems. For example, accurate performance

Manuscript received July 26, 2018; revised November 20, 2018; accepted December 24, 2018. This work was supported by the Advanced Grid Modeling Program at the U.S. Department of Energy Office of Electricity under Grant DE-OE0000875. Paper no. TPWRS-01154-2018. *(Corresponding author: Zhaoyu Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: yuanyx@iastate.edu; kavehdeh1@gmail.com; fbu@iastate.edu; wzy@iastate.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/TPWRS.2019.2893821

0885-8950 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

IEEE Proof

of Class I models depends on high penetration of real-time measurement units and availability of a sizable data history, which renders their practical implementation costly. On the other hand, Class II methods are generally based on the simplified assumption that the total daily energy consumption for each customer remains almost constant during a month. This assumption reduces the estimation accuracy. While in [9] a separation between weekday and weekend consumption data was introduced to alleviate this problem, this approach falls short of distinguishing load behavior in different individual days. In order to address these shortcomings, in this paper, a spectral clustering (SC)-based multi-timescale learning (MTSL) framework is proposed to estimate hourly load consumption for customers without SMs, using monthly billing data. In addition to identification of the typical daily load behaviors for observed customers [12], [13], the proposed method focuses on enhancing distribution network observability by inferring actual load characteristics of unmetered customers from those monitored with SMs. Unlike previous Class II methods that utilize the average daily consumption value to assess the daily load profile, the proposed model estimates the consumption values at different timescales to improve the load estimation performance. To achieve this, three stages are included in the load estimation framework: 1) Typical daily load profiles are classified and stored in a databank using a SC algorithm trained by the AMI dataset of *observed customers* (i.e., customers with SMs) [14]. 2) For each class of typical load behavior, a multi-layer MTSL model is developed, which can decompose the monthly consumption into different timescale components, such as weekly, daily, and hourly consumption. At each layer, a series of machine learning models are used to allocate energy consumption at slower timescale among faster timescale consumption variables. 3) Due to the absence of real-time data for unobserved customers without SMs, a branch current state estimation (BCSE)-aided method is proposed to identify their underlying typical daily consumption [15]. The residuals of BCSE are used to calculate the probability of all classes using a recursive Bayesian learning (RBL) approach [16]. The class with the highest probability is selected as the underlying typical load behavior for the unobserved customer. While this method is trained using SM data from observed distribution systems, it can be employed to estimate the hourly load data for a fully unobservable network without SMs. In [17] and [18], a conceptually-similar three-stage framework is provided to perform peak demand estimation for unmonitored low voltage (LV) substations using typical substation-level load profiles. However, our work pursues a distinct goal of inferring hourly demand for the unobserved customers at the grid-edge. The difficulty we face at the grid-edge, is the higher uncertainty of customer-level load, which makes the construction of pattern bank and demand inference challenging. Meanwhile, to monitor the system states, it is necessary to obtain the time-series customer pseudo load rather than the daily substation peak demand. Moreover, another challenging issue at the grid-edge is the unavailable context information of customers. Our multi-timescale three-stage customer demand inference model addresses these challenges by only relying on monthly billing data of unobserved customers, SM data of observed customers, and SCADA measurements. The proposed method has been tested using real utility data and compared with existing methods in the literature.

The rest of this paper is constructed as follows: Section II introduces the proposed observability enhancement framework. In Section III, a SC algorithm is utilized to build the consumption pattern bank for different types of customers. In Section IV, the MTSL method is presented. Section V formulates the BCSE-aided pattern identification approach. The numerical results are analyzed in Section VI. Section VII concludes the paper with major findings.

II. Introduction to Real Data and Proposed Observability Enhancement Framework

A. AMI Data Description

The available AMI data history contains several U.S. midwest utilities' hourly energy consumption data (kWh) for over 6000 customers. The data ranges from January 2015 to May 2018. While a few industrial consumers are included in the dataset, over 95% of customers are residential and commercial loads. The hourly data was initially processed to remove missing data caused by communication error. Then, the AMI dataset was divided into six separate subsets where each subset corresponds to weekday or weekend load profiles of residential, commercial and industrial customers.

B. Proposed Observability Enhancement Framework

The objective of this paper is to design a load estimation approach for fully or partially unobservable networks to avoid overmuch assumptions in the location/type of measurement units and availability of context information. Given that monthly billing data of consumers is generally available in all distribution systems, the data resource required for training the proposed load estimation approach consists of unobserved customers' monthly billing data and a limited number of AMI data from other observed networks. Extra available context information can also be added to improve the performance of the model but is not required. Different stages of the proposed observability enhancement framework are presented in Fig. 1.

- **Stage I - Consumption Pattern Bank:** Based on the six data subsets defined above, a SC algorithm is used to detect similarities in the diverse daily load profiles and define customer classes accordingly. As shown in Fig. 1, the results of clustering, $\{C_1, C_2, \ldots, C_M\}$, are stored in the specific consumption pattern bank according to the customer type, with each cluster representing a typical daily load profile. The pattern bank clustering results are stored and employed for the development of machine learning models (detailed in Section III).
- **Stage II - Multi-Timescale Consumption Inference:** A separate multi-layer MTSL model is trained for each class of customers using SM data of observed customers to convert the monthly billing data to hourly load values. In each MTSL model, machine learning algorithms are developed based on various pre-determined timescales. The customer

IEEE Proof

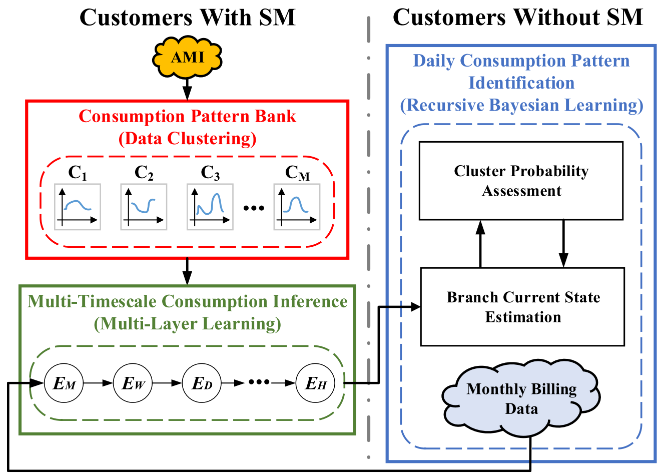


Fig. 1. Proposed observability enhancement framework.

consumption at these timescales are defined as monthly consumption E_M, weekly consumption E_W, daily consumption E_D, and hourly consumption E_H. The monthly data is regarded as the input for the first layer of the model and the hourly consumption variables appear in the output of the final layer. After the individual MTSL model of different classes are developed, the hourly estimation of unobserved customers are inferred by these models (detailed in Section IV).

- **Stage III - Consumption Pattern Identification:** In practice, the real hourly load of unobserved customers are unavailable *a priori* to determine the homologous daily load patterns. Hence, to assign a class from the daily pattern databank (Stage I) to unobserved customers, a BCSE-aided RBL method is proposed to identify these customers' underlying daily load profiles. Different daily profiles and their respective MTSL models are used for running BCSE over the target network for a period of time. The measurement residuals for each daily pattern are observed and utilized to make a connection between unobserved customers and their correct daily consumption patterns. Based on the observed residuals, a RBL method is employed to recursively assign a probability value to each typical daily consumption pattern for each unobserved customer. Then, the model with the highest probability is identified as the "correct" daily profile. The MTSL corresponding to the identified class for an unobserved customer is used to generate hourly pseudo measurements for that customer providing the redundancy to enhance the system observability (more details in Section V).

III. Proposed Clustering Algorithm

With the advent of AMI systems, typical daily load profile classification can be performed using different clustering algorithms, such as K-means, self-organizing maps, and hierarchical clustering [19]. In this paper, a graph theory-based clustering technique known as SC is utilized to distinguish the typical load profiles of observed customers and to create the typical consumption pattern bank. According to the properties of graph Laplacian, SC algorithm employs eigenvectors of graph matrices for data reconstruction. This reconstruction process enhances the cluster-properties in the data, so that clusters can be easily detected from the reconstruction datasets [20]. The improved cluster-properties of reconstructed datasets reduce the sensitivity of the clustering process to outliers [21]. Hence, the SC is robust and outperforms traditional clustering techniques, such as k-means, when tested on complex and unknown customer load shapes [22], [23]. In this paper, we apply automatic neighbor detection to avoid error from manual parameter selection and the main steps of SC are listed as follows [14]:

- **Step I:** As a graph theoretic clustering approach, SC algorithm transforms AMI dataset into a similarity graph $G = (V, E)$, which consists of a set of vertices V and a set of edges E connecting different vertices. For our problem, vertices V are constructed by using the average daily load profile of observed customers. Hence, V_i is the average load consumption of i'th customer: $V_i = [\overline{E^i_{H1}}, \ldots, \overline{E^i_{H24}}]$, where $\overline{E^i_{Hj}}$ indicates the average load value at the j'th hour of the i'th customer. The average hourly load profile is computed by $\overline{E^i_{Hj}} = \frac{1}{N_d}\sum_{d=1}^{N_d} E^i_{Hj}(d)$, where N_d is the total number of recorded days in the training set. Two vertices are connected if the corresponding pair-wise similarity is non-zero. In this paper, a technique is utilized for constructing fully-connected graphs, in which vertex V_i is connected to all vertices that have positive similarity with V_i. The goal of similarity graph is to model local neighborhood relations between data points. The value of similarity relies on a scaling parameter α that controls how rapidly the similarity weights, W_{ij}, fall off with the distance between vertices. Note that the *distance* between vertices a and b is defined as $||a - b||$ [20]. Instead of using a single α, we calculate a local α_i for each vertex V_i that allows self-tuning of the point-to-point distances, as $\alpha_i = ||V_i - V_K||$, where V_K is the K'th neighbor of vertex V_i.
- **Step II:** Based on the local scaling parameter α_i, the weighted adjacency matrix of the graph $W = (w_{i,j})_{i,j=1,\ldots,n}$ is developed. We have adopted the Gaussian kernel function to build the adjacency matrix W as follows:

$$w_{i,j} = exp\left(\frac{-||V_i - V_j||^2}{\alpha_i \alpha_j}\right) \tag{1}$$

- **Step III:** After the weighted adjacency matrix is built, SC converts the clustering process to a graph partitioning problem, which divides a graph into k disjoint sets of vertices by removing edges connecting each two groups. When the edges between different sets have low weight and the edges within a set have high weight, a satisfactory partition of the graph is obtained [22]. Hence, the objective function of graph partitioning is to maximize both the dissimilarity between the different clusters and the total

IEEE Proof

similarity within each cluster [24]:

$$N(G) = \min_{A_1,\ldots,A_\eta} \sum_{i=1}^{\eta} \frac{c(A_i, \overline{A_i})}{d(A_i)} \tag{2}$$

where, η is the number of vertices, A_i is a subset belonging to V, $c(A_i, \overline{A_i})$ is the sum of the weights between vertices in A_i and vertices in the rest of the subsets, $d(A_i)$ is the sum of the weights of vertices in A_i. It was proved in [20] that the minimum of $N(G)$ is obtained at the second smallest eigenvector of the Laplacian matrix. Graph Laplacian matrix is the main element of the SC algorithm and constructed using the adjacency matrix W and a diagonal matrix D whose (i,i)'th element is the sum of W's i'th row. The normalized graph Laplacian is given by [25]:

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{3}$$

- **Step IV:** When the associated Laplacian matrix $L \in \mathbb{R}^{n\times n}$ has been constructed using the similarity matrix W of vertex V_i, we compute the eigenvector $[y_1, y_2, \ldots, y_n]$ of the Laplacian matrix and pick the eigenvectors corresponding to the k smallest eigenvalues, where the range of k is $n \geq k \geq 2$. The first k eigenvectors are extracted to build a new matrix $Y \in \mathbb{R}^{n\times k}$. Due to the properties of the graph Laplacians, the vertex V_i is represented by the i'th row of the Y matrix. This change of representation enhances the cluster-properties in the data and a simple clustering algorithm is able to detect the clusters in the reconstructed data [22]. In this paper, we use the k-means algorithm to obtain the k corresponding clusters for the original vertex, V_i. It is feasible to utilized other techniques, such as the hyperplanes and advanced post-processing of the eigenvectors, to replace the k-means method to extract the final solution in this step [22].
- **Step V:** To find the best partitioning, the Davies-Bouldin validation index (DBI) is applied to calibrate the SC algorithm by measuring the ratio of within-cluster and between-cluster similarities [12]. Step IV is repeated with different k values, and corresponding DBI values for each k are recorded. The value of k for which DBI is minimized is chosen as the optimal number of clusters [26]. This process is applied to the rest of the data subsets to determine the number of typical load profiles.

IV. Inference of Hourly Energy Consumption

A MTSL method is assigned and trained for each typical load profile using the available data in the pattern bank defined in Section III, to map monthly consumption data to hourly load for customers belonging to each class. While hourly load variations cannot be directly observed at the monthly level, a multi-layer structure, where each layer corresponds to the total consumption at different timescales, is able to make this connection between monthly and hourly data with good accuracy. Hence, the MTSL is constructed in a way to keep a high correlation level between inputs-outputs of different layers to maintain layer-wise estimation accuracy. In order to identify variables with high correlation coefficient levels to design the structure of the MTSL, a

TABLE I
Statistical Multi-Timescale Consumption Analysis

Layer	Correlation	Industrial	Commercial	Residential
Layer I	$\rho(E_M, E_W)$	0.9744	0.9921	0.9613
	$\rho(E_W, E_W)$	0.9600	0.9843	0.9309
Layer II	$\rho(E_W, E_{D_w})$	0.9764	0.9875	0.9400
	$\rho(E_{D_w}, E_{D_w})$	0.9677	0.9862	0.8983
	$\rho(E_W, E_{D_{nw}})$	0.9234	0.9500	0.9281
	$\rho(E_{D_{nw}}, E_{D_{nw}})$	0.9241	0.9771	0.8871
Layer III	$\rho(E_{D_w}, E_{H_w})$	0.9498	0.9429	0.7747
	$\rho(E_{H_w}, E_{H_w})$	0.9838	0.9793	0.7882
	$\rho(E_{D_{nw}}, E_{H_{nw}})$	0.9573	0.9667	0.7728
	$\rho(E_{H_{nw}}, E_{H_{nw}})$	0.9881	0.9833	0.7960

basic statistical analysis was performed on the AMI dataset, as shown in Table I. The consumption levels at different timescales are defined as, monthly consumption E_M, weekly consumption E_W, weekday consumption E_{D_w}, weekend consumption $E_{D_{nw}}$, weekday hourly consumption E_{H_w}, and weekend hourly consumption $E_{H_{nw}}$, and obtained using hourly SM data history. For different types of customers, the correlation values are shown in Table I and determined as follows:

$$\rho(X,Y) = \left| \frac{\sigma^2_{X,Y}}{\sigma_X \sigma_Y} \right| \tag{4}$$

where, X and Y are the consumption levels of observed customers at specific timescales, such as monthly or weekly consumption. $\sigma^2_{X,Y}$ is the covariance of X and Y, and σ_X defines the standard deviations of the variable. Using the correlation analysis, a three-layer structure is developed for each type of customer and typical load behavior stored in the pattern bank, as shown in Fig. 2. In this figure, Layer I converts total monthly consumption, E_M, to the set of weekly consumption values $E_W = \{E_{W1}, \ldots, E_{W4}\}$ using ANNs connected in series. To capture the temporal correlation between consumption at consecutive weeks, each week's estimated consumption is also fed to the next ANN corresponding to the following week's consumption. This idea is shown in (5) and generalized to all the layers of MTSL, as demonstrated in Fig. 2:

$$E_{Wi} = ANN\left(E_M, E_{W(i-1)}\right) \tag{5}$$

The output of Layer I forms the weekly training set that becomes the input of Layer II. This layer converts weekly consumption, E_W, to the set of daily consumption $E_D = \{E_{D1}, \ldots, E_{D7}\}$ by various ANNs. Based on the distinct customer behavior on weekdays and weekends, Layer III is trained to map the total daily consumption to hourly consumption $E_H = \{E_{H1}, \ldots, E_{H24}\}$. In the proposed model, the Levenberg-Marquardt (LM) backpropagation method is used to update the network weight and bias variables [27]. The LM algorithm is derived from Newton's method to minimize sum-of-square error functions [28]. Compared to backpropagation algorithms with a constant learning rate, LM can automatically adjust the learning

IEEE Proof

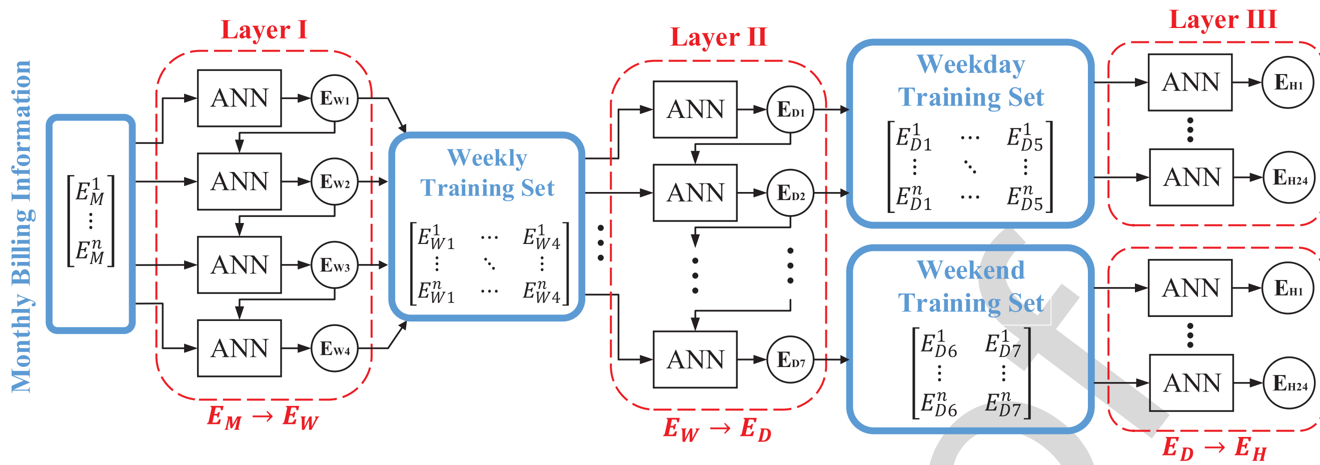


Fig. 2. Multi-timescale learning structure.

rate in the direction of gradient using the Hessian matrix, which significantly increases the training speed [29], [30]. The training objective function (F) and the update equation of LM can be written as:

$$\min_b F(b) = \sum_{i=1}^{Q} v_i^2(b) = v^T(b)v(b) \tag{6}$$

$$\Delta b_l = -\left[J^T(b_l)J(b_l) + \mu_l I\right]^{-1} J^T(b_l)v(b_l) \tag{7}$$

where, μ_l is the combination parameter at iteration l, b is the set of learning parameters, J is the training objective function's Jacobian, I is the identify matrix, v is the error vector, T is the matrix transposition operation, and Δb_l defines the learning parameter updates at each iteration. In each iteration, the value of μ_l is updated based on the change of approximated performance index $F(b)$. If a smaller value is obtained, the μ_l is divided by some factor $\vartheta > 1$. Otherwise, μ_l is multiplied by ϑ for the next iteration.

For each ANN, the dataset is randomly divided into three separate subsets for training (70% of the total data), validation (15% of the total data), and testing (15% of the total data). To calibrate the hyper-parameters of each ANN, we utilize the grid search methods to find the optimal sets of four important parameters of LM: the number of hidden layer, the number of neurons, the value of increase factor ϑ and the value of decrease factor $\frac{1}{\vartheta}$ [31]. As a multi-layer structure with a high number of learning parameters, the *overfitting* problem poses a critical risk against reliability of the learned model. Overfitting is a result of model over-flexibility which occurs when the model shows low bias but high variance [32]. In order to overcome this problem, we have adopted two approaches in this paper: 1) *Early stopping mechanism,* in which the training process is terminated as soon as the validation error starts to increase [33]. 2) *Noise injection,* which improves the robustness of ANNs by injecting small noise to the AMI training sets [34].

V. Proposed Method for Pattern Identification

In the proposed approach, various MSTL models are assigned to typical consumption patterns. In practice, monthly billing data alone is not enough to determine the typical load profiles of unobserved customers. The pervasive real-time data source in distribution systems is a limited number of feeder-level measurements, such as SCADA voltage and current measurements. In order to identify and allocate the corresponding daily pattern and related MSTL to unobserved customers using only feeder-level measurements, a BCSE-aided RBL method is proposed [16]. This learning algorithm computes the probability of each typical load pattern for an unobserved customer using the residuals of a BCSE algorithm [15]. Based on the probability values, the most probable class is chosen as the correct underlying profile for unobserved customer.

A. BCSE

A BCSE algorithm is tailored for real-time monitoring of distribution systems [15] [35]. Compared to traditional state estimation methods that use node voltages as system states, BCSE is shown to improve the computational efficiency and memory requirements by adopting branch currents as state variables. In general, the Weighted Least Square (WLS) algorithm is widely-used to solve the BCSE problem to obtain an estimation of system nodes [36]. The objective function of WLS is defined as follows:

$$\min_x J = (z - h(x))^T \Sigma (z - h(x)) \tag{8}$$

where, z is the measurement vector, x is the state vector, i.e., $x = [I_r, I_x]$ with I_r and I_x representing the branch currents' real part and branch currents' imaginary part, h is the nonlinear measurement function associated with measurement z. The residual vector of BCSE is defined as the difference between the real measurements with estimated values, $r = z - h(x)$, and Σ denotes the weight matrix that represents the accuracy of measurements. In general, the variance of the measurement error, φ^2, is used to build Σ, as $\Sigma = diag\{\varphi_1^{-2}, \ldots, \varphi_s^{-2}\}$, where s

IEEE Proof

represents the cardinality of z [37]. The Gauss-Newton method is adopted to solve this non-convex optimization problem [15]. The basic idea of Gauss-Newton method is to find a solution for $\nabla_x J = 0$, where $\nabla_x J$ denotes the gradient of J with respect to state variables. The iterative processes of the algorithm are as follows:

$$G(x) = H^T(x)\Sigma H(x) \tag{9}$$

$$[G(x^m)]\Delta x^m = H^T(x^m)\Sigma(z - h(x^m)) \tag{10}$$

$$x^{m+1} = x^m + \Delta x^m \tag{11}$$

where, H is the Jacobian matirx of the measurement function $h(x)$, G is the gain matrix, and m is the iteration number.

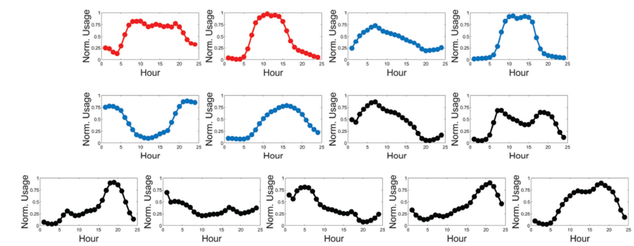
B. Load Pattern Assignment by RBL

To identify the underlying daily consumption pattern for unobserved customers, the following steps are performed:

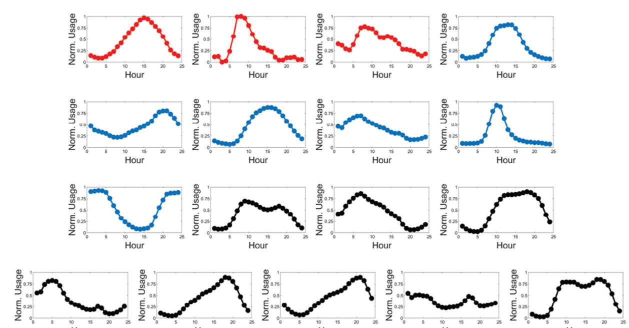
- **Stage I:** Select a class, denoted as i, from the daily consumption pattern bank, for unobserved customer j.
- **Stage II:** Use the MSTL of the selected class to generate hourly pseudo load values from the customer's monthly billing data.
- **Stage III:** Run the BCSE using the generated pseudo load values. Observe the residuals. The residuals of each estimator can be obtained by comparing the real measurements with estimated values.
- **Stage IV:** Define probability $p_{i,j}$ as: "the probability that class i is the correct average daily consumption profile for customer j." The initial value of $p_{i,j}$ is defined as $\frac{1}{N}$ for iteration count 0, where N is the number of MSTL models for a specific customer type [16]. Applying the Bayes theorem and assuming a Gaussian distribution for measurement error, a recursive expression for updating this probability over time is obtained as follows [38]:

$$p_{i,j}^o = \frac{\exp(-\frac{1}{2} r_{i,j}^{o^T} \cdot \Phi \cdot r_{i,j}^o) p_{i,j}^{o-1}}{\sum_{t=1}^N \exp(-\frac{1}{2} r_{t,j}^{o^T} \cdot \Phi \cdot r_{t,j}^o) p_{t,j}^{o-1}} \tag{12}$$

  where, o is the iteration count, $r_{i,j}^o$ is the residual vector of the i'th class with respect to j'th customer and is computed by the corresponding state and real measurement vectors $r_{i,j}^o = z - h(x_i^o)$, Φ is a diagonal matrix that represents the variances corresponding to the residual components $\Phi = diag\{\sigma_{r_{i,j},R}^2, \sigma_{r_{i,j},I}^2\}$ to increase the speed of convergence, where $\sigma_{r_{i,j},R}^2$ is the variance of the branch current real part residual and $\sigma_{r_{i,j},I}^2$ is the variance of the branch current imaginary part residual.
- **Stage V:** Go back to Stage I.
- **Stage VI:** Identify the underlying daily load profile for the unobserved customer, i^*, as the most probable class: $i^* = \text{argmax}_i\, p_i^j$.
- **Stage VII:** Repeat the above process for all unobserved customers until the average daily load profiles of all customers are identified.
- **Stage VIII:** Perform online BCSE for real-time system monitoring using MTSL-based pseudo hourly load estimations obtained from the assigned classes to unobserved customers.



(a) Industrial (red), commercial (blue), and residential (black) weekday typical load pattern



(b) Industrial (red), commercial (blue), and residential (black) weekend typical load pattern

Fig. 3. Consumption pattern bank for industrial, commercial, and residential customers on weekday and weekend.

The main advantage of the RBL is exponential rejection of the wrong load patterns and low computational complexity which is advantageous in large distribution systems [16].

VI. Numerical Results

The proposed observability enhancement framework is tested for unobserved customers on a real distribution feeder, shown in Fig. 4. This feeder contains three types of loads: industrial (3%), commercial (20%), and residential (77%) loads. The proposed method is compared with two existing load estimation approaches adopted from [9] and [11], in terms of accuracy.

A. Calibration Performance

To calibrate the parameters of SC and ANN, the DBI index and grid search are utilized to find the optimal parameters. For the SC method, the optimal number of cluster, k, is obtained based on the minimum DBI value, as shown in Fig. 7. For the calibration of ANN, the the optimal hyper-parameter set is decided by the grid search method [31]. Due to page limit, we have presented a sample grid search calibration result for one ANN in Fig. 7.

B. SC Algorithm Performance

Based on the AMI dataset, the SC algorithm is utilized to classify different load shapes and to create the consumption pattern banks. Fig. 3 shows typical load patterns for different types of customers for weekdays and weekends. As shown in

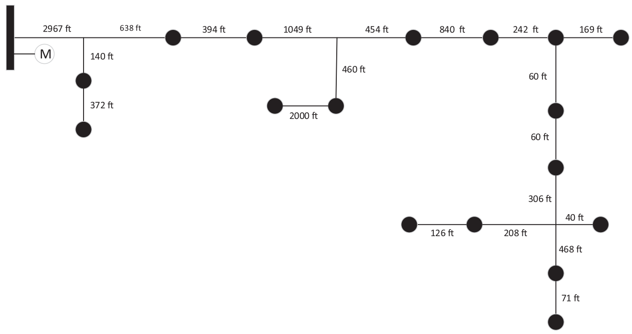


Fig. 4. A 18-node real utility feeder case.

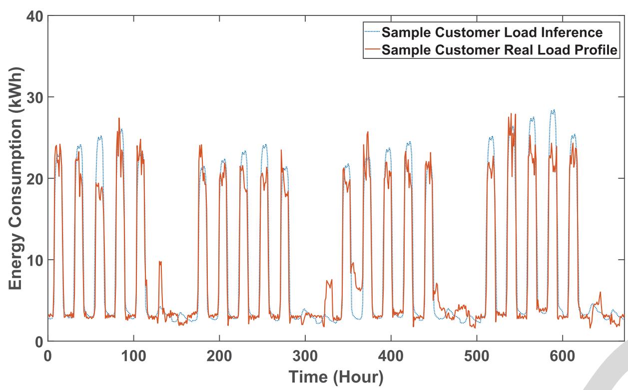


Fig. 5. Comparison of hourly load inference with real load profile.

Fig. 3, the numbers of typical load profiles in weekdays are normally smaller than that of weekends. Compared to the diverse activities in weekends, customers have relatively few normative load behaviors in weekdays. Also, as expected, the residential customers have more load patterns than industrial and commercial customers due to the higher variation of residential load behaviors.

C. Pseudo Measurement Generation Performance

After consumption pattern banks have been developed from AMI data of observed systems, the muti-layer MSTL models are trained and tested on the feeder shown in Fig. 4. In this case, the test feeder is considered to be a fully unobserved network in which no customer is equipped with SMs. To reduce the error of the learning model, the MTSL method has been tested over 12-month load data. Fig. 5 shows the comparison between hourly load inference of one sample customer, obtained from monthly billing data, and real load profile during that month. As can be seen, the pseudo hourly load samples are able to accurately track the customer's real consumption. Fig. 6 presents the accuracy comparison of load estimation for different types of customers. The monthly data of test customers are used as the input of all MSTL models. The *goodness-of-fit* measure, R^2, is used to assess the accuracy of the result, with $R^2 = 1$ indicating a perfect fit. The R^2 values are used to measure the accuracy of MTSLs corresponding to correct and incorrect daily pattern consumption classes for all customers. The R^2 is computed by

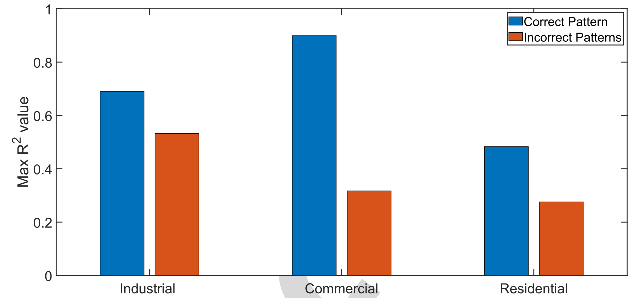


Fig. 6. Customer level load estimation result.

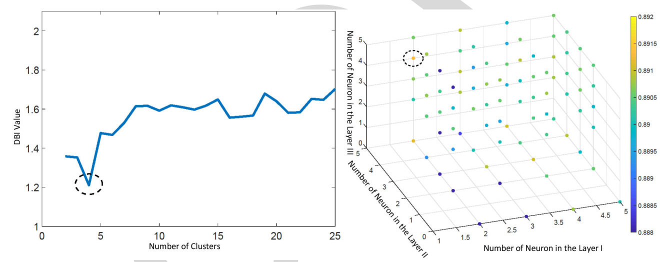


Fig. 7. Calibration result of SC (left) and ANN (right).

the total sum of squares of estimation error and deviation from mean. The equation is given as the follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{J} (\tau_i - f_i)^2}{\sum_{i=1}^{J} (\tau_i - \overline{\tau})^2} \tag{13}$$

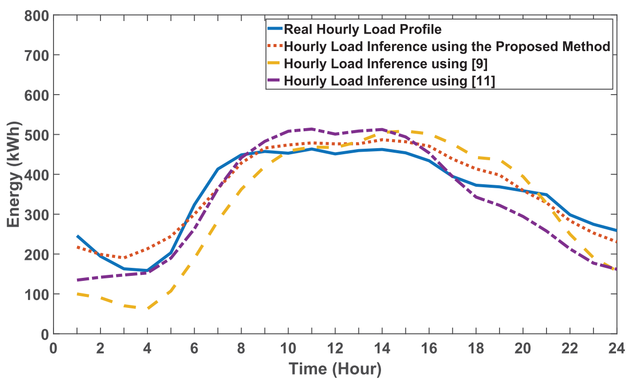
where, f_i is the estimated value, τ is the observed data and $\overline{\tau}$ is the mean of the observed data. As expected, the MTSL load estimation model corresponding to the correct underlying consumption class for the customers has a better accuracy, compared to the incorrect one. This further supports the correct functionality of RBL, as described in the next subsection. Also, as shown in Fig. 6, for industrial and commercial customers, the learning model yields more accurate estimations compared to the residential customers due to lower consumption volatility. In contrast, for residential customers, the diversity and complexity of human activities lead to less accurate estimations.

Fig. 8 shows the feeder-level daily load estimation results (in weekdays and weekends) averaged over a total of 15 months for our proposed learning model and two existing methods in the literature [9] [11]. The Mean Absolute Percentage Error (MAPE) criterion is utilized to evaluate the accuracy of estimation methods:

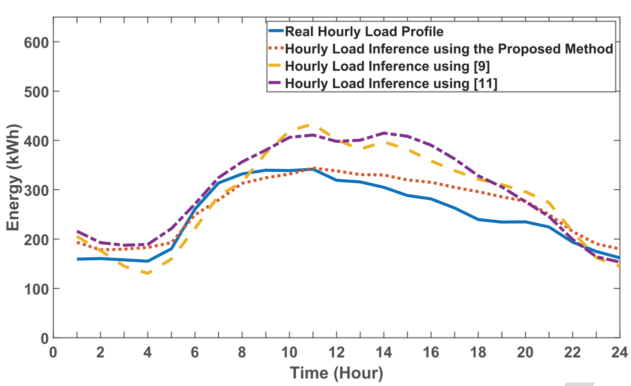
$$M = \frac{100\%}{n_s} \sum_{t=1}^{n_s} \left| \frac{A(t) - E\{A(t)\}}{A(t)} \right| \tag{14}$$

where, A is the actual load value and $E\{\cdot\}$ is the mean operator. As is demonstrated in these figures, the estimation MAPE values for the proposed method are $\{7.40\%, 10.02\%\}$ for weekdays and weekends, respectively. On the other hand, the proposed methods in [9] and [11] show average MAPE of $\{19.47\%, 20.32\%\}$ and $\{13.79\%, 21.16\%\}$ over the test set. Hence, based on this

IEEE Proof



(a) Sample feeder average daily load inference results in weekday



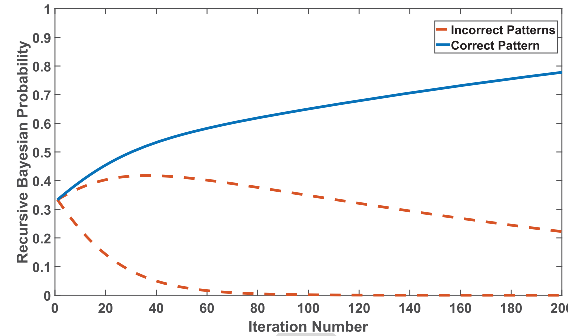
(b) Sample feeder average daily load inference results in weekend

Fig. 8. Comparison of load inference results.

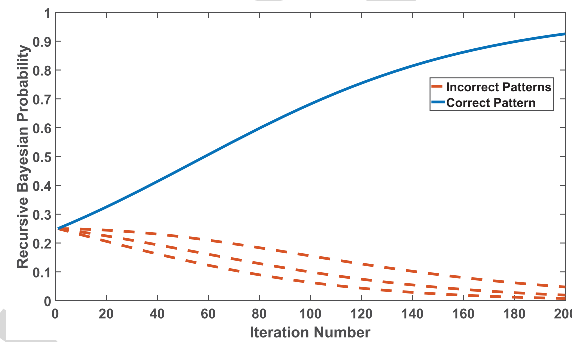
AMI dataset and the test feeder, the proposed method shows a better accuracy for hourly load inference compared to the previous works.

D. Load Pattern Identification

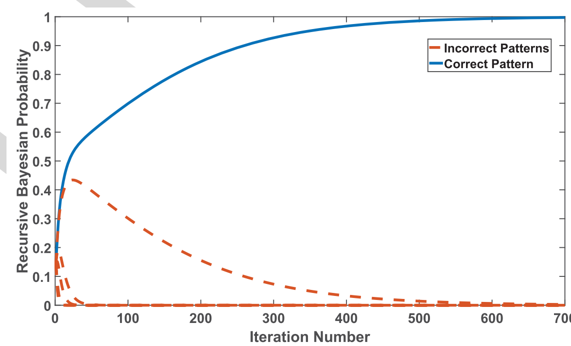
The performance of the BCSE-aided pattern identification scheme was tested on three cases of different types of customers, corresponding to industrial, commercial, and residential loads. A Phasor Measurement Unit (PMU) was placed at the main bus of the test feeder to provide the real measurement value for BCSE. Pseudo hourly load estimations were extracted from unobserved customers' monthly billing data, for different candidate daily consumption profiles in the databank. According to the residuals, the graphs in Fig. 9 show the probabilities assigned by the RBL algorithm to the correct and incorrect load patterns available in the typical daily load profile bank. Over the iterations, one MSTL model has the asymptotic probability close to one while others have almost 0 probabilities. Based on the previous work [16], the model with the highest probability is identified as the target model. As is demonstrated in Fig. 9, the proposed algorithm is effective since it successfully identifies the MTSL model corresponding to the correct latent daily consumption pattern, by assigning the highest probability value to it for all types of customers.



(a) Sample industrial customer identification



(b) Sample commercial customer identification

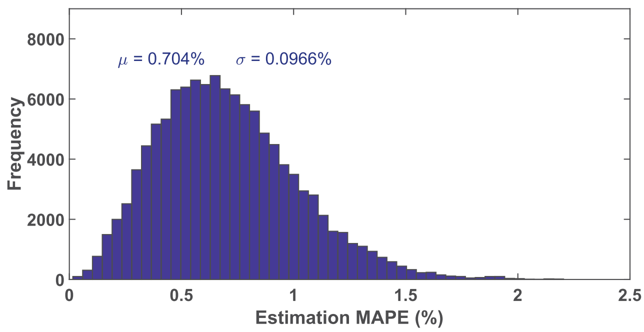


(c) Sample residential customer identification

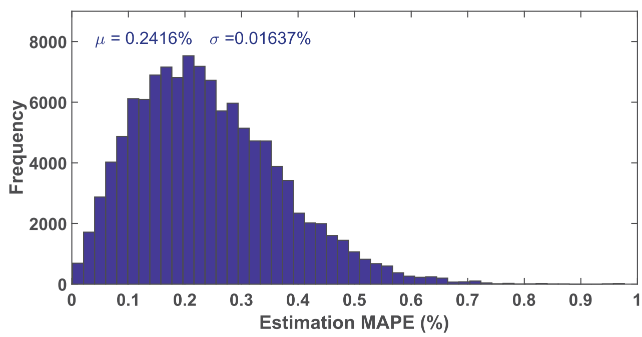
Fig. 9. Performance of BCSE-aided RBL daily profile identification method for three types of customers.

E. State Estimation Performance

After hourly pseudo measurement samples are generated for every unobserved customer using the proposed method, BCSE can be performed in real-time over the test feeder given the introduced data-driven redundancy. The error distribution of real-time state estimation is shown in Fig. 10 for voltage magnitude and phase components. As is demonstrated in the figure, based on the proposed load estimation approach, BCSE can obtain system state estimation with magnitude and phase angle estimation mean errors of 0.70% and 0.24%, respectively. In the previous work [35], the mean errors of voltage magnitude and phase angle are around 0.73% and 0.36%, respectively in the BCSE algorithm with 20% maximum error for pseudo measurements. Hence, by comparison, our BCSE and machine learning framework shows a comparably valid performance.



(a) Voltage magnitude component error



(b) Voltage phase component error

Fig. 10. BCSE-based state estimation performance using the proposed load inference model.

VII. Conclusion

In this paper, we have presented a data-driven method for load estimation to improve the observability of distribution systems without AMI. The proposed method is able to extract hourly load estimations from monthly billing data for all types of customers, including residential, commercial, and industrial. Moreover, this approach can identify the average daily load pattern of unobserved customers using a BCSE-aided probabilistic learning method. The proposed method is successfully validated on a real utility feeder with real SM data and has been able to improve the performances of existing methods in the literature.

References

[1] Office of Electricity Delivery and Energy Reliability, "Advanced metering infrastructure," Feb. 2008. [Online]. Available: https://www.energy.gov/sites/prod/files/2016/12/f34/AMI ort_09-26-16.pdf

[2] Energy Information Administration, "Advanced metering count by technology type," 2017. [Online]. Available: https://www.eia.gov/electricity/annual/html/epa_10_10.html

[3] O. Chilard, S. Grenard, O. Devaux, and L. de Alvaro Garcia, "Distribution state estimation based on voltage state variables : Assessment of results and limitations," in *Proc. 20th Int. Conf. Exhib. Electricity Distrib. Part 1*, Jun. 2009, pp. 1–4.

[4] A. Al-Wakeel, J. Wu, and N. Jenkins, "k-means based load estimation of domestic smart meter measurements," *Appl. Energy*, vol. 194, no. 1, pp. 333–342, May 2017.

[5] Y. Li and P. J. Wolfs, "A hybrid model for residential loads in a distribution system with high PV penetration," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3372–3379, Aug. 2013.

[6] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Jrventausta,"Enhanced load profiling for residential network customers," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 88–96, Feb. 2014.

[7] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 548–555, May 2005.

[8] E. Manitsas, R. Singh, B. C. Pal, and G. Strbac, "Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 1888–1896, Nov. 2012.

[9] Y. R. Gahrooei, A. Khodabakhshian, and R. A. Hooshmand, "A new pseudo load profile determination approach in low voltage distribution networks," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 463–472, Jan. 2018.

[10] J. A. Jardini, C. M. V. Tahan, M. R. Gouvea, S. U. Ahn, and F. M. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," *IEEE Trans. Power Del.*, vol. 15, no. 1, pp. 375–380, Jan. 2000.

[11] D. T. Nguyen, "Modeling load uncertainty in distribution network monitoring," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2321–2328, Sep. 2015.

[12] G. J. Tsekouras, P. B. Kotoulas, C. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Elect. Power Syst. Res.*, vol. 78, pp. 1494–1510, Jun. 2008.

[13] G. J. Tsekouras, I. Hatzilau, and J. Prousalidis, "A new pattern recognition methodology for classification of load profiles for ships electric consumers," *J. Marine Eng. Technol.*, no. A14, pp. 45–58, 2009. Q3

[14] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 1601–1608.

[15] M. E. Baran and A. W. Kelley, "A branch-current-based state estimation method for distribution systems," *IEEE Trans. Power Syst.*, vol. 10, no. 1, pp. 483–491, Feb. 1995.

[16] R. Singh, E. Manitsas, B. C. Pal, and G. Strbac, "A recursive Bayesian approach for identification of network configuration changes in distribution system state estimation," *IEEE Trans. Power Syst.*, vol. 25, no. 3, pp. 1329–1336, Aug. 2010.

[17] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templatespart I: Substation clustering and classification," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3036–3044, Nov. 2015.

[18] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templatespart II: Peak load estimation by clusterwise regression," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3045–3052, Nov. 2015.

[19] S. M. Bidoki, N. Mahmoudi-Kohan, M. H. Sadreddini, M. Z. Jahromi, and M. P. Moghaddam, "Evaluating different clustering techniques for electricity customer classification," in *Proc. IEEE PES Transmiss. Distrib. Conf. Expo.*, 2010, pp. 1–5.

[20] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *in Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[21] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, Jun. 2015.

[22] U. Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Mar. 2007.

[23] D. Vercamer, B. Steurtewagen, D. V. den Poel, and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3693–3701, Sep. 2016.

[24] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1944–1957, Nov. 2007.

[25] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 1997.

[26] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. Energy*, vol. 141, pp. 190–199, Mar. 2015.

[27] S. Sapna, A.Tamilarasi, and P. Kumar, "Backpropagation learning algorithm based on Levenberg–Marquardt algorithm," *Comput. Sci. Inf. Technol.*, pp. 393–398, 2012.

[28] C. L. *et al.*, "Levenberg–Marquardt backpropagation training of multilayer neural networks for state estimation of a safety-critical cyber-physical system," *IEEE Trans. Ind. Inform.*, vol. 14, no. 8, pp. 3436–3446, Aug. 2018.

[29] B. M. Wilamowski and H. Yu, "Improved computation for Levenberg–Marquardt training," *IEEE Trans. Neural Netw.*, vol. 21, no. 6, pp. 930–937, Jun. 2010.

[30] N. Zhang and P. K. Behera, "Solar radiation prediction based on recurrent neural networks trained by levenberg-marquardt backpropagation learning algorithm," in *Proc. IEEE PES Innovative Smart Grid Technologies*, 2012, Jan. pp. 1–7.

[31] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[32] I. Bilbao and J. Bilbao, "Overfitting problem and the over-training in the era of data: Particularly for artificial neural networks," in *Proc. 8th Int. Conf. Intell. Comput. Inf. Syst.*, Dec. 2017, pp. 173–177.

Q4 [33] C. Doan and S. Liong, "Generalization for multilayer neural network Bayesian regularization or early stopping," in *Proc. 2nd Conf. Asia Pac. Assoc. Hydrol. Water Resour.*, Jan. 2004.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[35] H. Wang and N. N. Schulz, "A revised branch current-based distribution system state estimation algorithm and meter placement impact," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 207–213, Feb. 2004.

[36] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. New York, NY, USA: Marcel Dekker, 2004.

Q5 [37] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan, and F. Bu, "A survey on state estimation techniques and challenges in smart distribution systems," *IEEE Trans. Smart Grid*, to be published.

[38] S. Wang and Y. Zhao, "Online Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 663–677, Sep. 2001.

Yuxuan Yuan (S'18) received the B.S. degree in electrical and computer engineering from Iowa State University, Ames, IA, USA, in 2017, where he is currently working toward the Ph.D. degree. His research interests include distribution system state estimation, synthetic networks, data analytics, and machine learning.

Kaveh Dehghanpour (S'14–M'17) received the B.Sc. and M.S. degrees in electrical and computer engineering from the University of Tehran, Tehran, Iran , in 2011 and 2013, respectively, and the Ph.D. degree in electrical engineering from Montana State University, Bozeman, MT, USA, in 2017. He is currently a Postdoctoral Research Associate with Iowa State University, Ames, IA, USA. His research interests include application of machine learning and data-driven techniques in power system monitoring and control.

Fankun Bu (S'18) received the B.S. and M.S. degrees from North China Electric Power University, Baoding, China, in 2008 and 2013, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From 2008 to 2010, he worked as a Commissioning Engineer with NARI Technology Co., Ltd., Nanjing, China. From 2013 to 2017, he worked as an Electrical Engineer with the State Grid Corporation of China at Jiangsu, Nanjing, China. His research interests include load modeling, load forecasting, distribution system estimation, machine learning, and power system relaying.

Zhaoyu Wang (S'13–M'15) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, Shanghai, China, in 2009 and 2012, respectively, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2012 and 2015, respectively. He is the Harpole-Pentair Assistant Professor with Iowa State University, Ames, IA, USA. He was a Research Aid at Argonne National Laboratory in 2013 and an Electrical Engineer Intern with Corning Inc. in 2014. His research interests include power distribution systems, microgrids, renewable integration, power system resilience, and power system modeling. He is the Principal Investigator for a multitude of projects focused on these topics and funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Energy Center. He was the recipient of the IEEE PES General Meeting Best Paper Award in 2017 and the IEEE Industrial Application Society Prize Paper Award in 2016. He is the Secretary of the IEEE Power and Energy Society Award Subcommittee. He is an editor for the IEEE TRANSACTIONS ON SMART GRID and IEEE PES LETTERS.

IEEE Proof