# A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand

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

***Abstract*—Advanced metering infrastructure (AMI) enables utilities to obtain granular energy consumption data, which offers a unique opportunity to design customer segmentation strategies based on their impact on various operational metrics in distribution grids. However, performing utility-scale segmentation for *unobservable* customers with only monthly billing information, remains a challenging problem. To address this challenge, we propose a new metric, the coincident monthly peak contribution (CMPC), that quantifies the contribution of individual customers to system peak demand. Furthermore, a novel multi-state machine learning-based segmentation method is developed that estimates CMPC for customers without smart meters (SMs): first, a clustering technique is used to build a databank containing typical daily load patterns in different seasons using the SM data of observable customers. Next, to associate unobservable customers with the discovered typical load profiles, a classification approach is leveraged to compute the likelihood of daily consumption patterns for different unobservable households. In the third stage, a weighted clusterwise regression (WCR) model is utilized to estimate the CMPC of unobservable customers using their monthly billing data and the outcomes of the classification module. The proposed segmentation methodology has been tested and verified using real utility data.**

***Index Terms*—Customer segmentation, peak load contribution, observability, machine learning.**

## I. Introduction

Advent of Advanced metering infrastructure (AMI) has facilitated a deeper understanding of customer behaviors in low-voltage networks for distribution system operators. Individual customers' demand consumption can be recorded by smart meters (SMs) with high temporal resolution, which enables developing novel data-centric grid operation mechanisms. One of these mechanisms is utility-scale customer segmentation [1], which is extremely useful in enhancing system operation and management by intelligently targeting customers for peak shaving programs, AMI investment, and retail price/incentive design. This will help utilities under strict financial constraints to optimize their investment portfolio. However, for small-to-medium utilities, a key barrier against investigating an efficient customer segmentation is the absence of real-time measurements due to financial limitations [2]. Currently, more than half of all U.S. electricity customer accounts do not have SMs to record their detailed consumption behavior [3].

Several papers have focused on developing customer segmentation strategies using SM data. One of the most common approaches is to leverage clustering techniques for identifying typical load profiles [4]–[6]. In [4], principal component analysis (PCA) is performed to extract the dominant features within customer consumption data and then k-means algorithm is employed to classify consumers. In [5], a finite mixture model-based clustering is presented to obtain distinct behavioral groups. In [6], a C-vine copulas-based clustering framework is proposed to carry out consumer categorization. However, the typical load profile extraction alone is insufficient to assess customers‘ impacts on system peak demand, which limits utilities' ability to target suitable customers for reducing the operation costs.

Apart from typical load profiles, several customer segmentation methodologies have been developed based on the feature characterization and extraction [7]–[10]. In [7], residential customers are ranked using their appliance energy efficiency to reduce building energy consumption. In [8], the entropy of household power demand is used to evaluate the variability of consumption behavior, which is considered to be a key component in peak shaving program targeting and customer engagement. In [9], a customer's marginal contribution to system cost is obtained using daily demand profiles. In [10], a four-stage data-driven probabilistic method is proposed to estimate the coincident peak demand estimation of new customers for designing new systems. Compared to the clustering approaches, these methods directly quantify customer-level features from SM data and use them to determine the segmentation strategies. Nevertheless, the previously-proposed metrics fall short of considering customers' impact on system peak demand, which is a major problem considering that continuous growth in system peak load raises the possibility of power failure and increases the marginal cost of supply [11]. Furthermore, previous works have only focused on observable customers.

In order to address these shortcomings, this paper proposes a new metric for customer segmentation, which is denoted as coincident monthly peak contribution (CMPC). CMPC is defined

Q1 Manuscript received October 5, 2019; revised January 17, 2020; accepted March 7, 2020. This work was supported in part by the Advanced Grid Modeling Program at the U.S. Department of Energy Office of Electricity under Grant DE-OE0000875, and in part by the National Science Foundation under Grant ECCS 1929975. Paper no. TPWRS-01502-2019. *(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 article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2020.2979943

as the ratio of individual customer's demand during system daily peak load time over the real-time total system peak demand in a course of a month. Compared with conventional coincident peak demand metrics, which quantify the peak consumption levels of multiple customers based on their empirical diversified maximum demand [10], the proposed CMPC focuses on the impact of individual customer and conveys information on how individual customer's peak time differs from the system's peak demand time. Based on the definition of CMPC, we develop a multi-stage machine learning-based customer segmentation strategy that estimates CMPCs of unobservable customers using only their monthly billing information. The developed method consists of three modules: 1) Using a graph theoretic clustering, a seasonal typical load pattern bank is constructed to classify various customer consumption behaviors. 2) To connect unobservable customers to the seasonal databank, a multinomial classification model is presented which identifies typical load profiles of customers without SMs. 3) According to the outcome of the classification module, a weighted clusterwise regression (WCR) model is trained to map the unobservable customers' monthly energy consumption data to CMPC values. Utilizing our segmentation method, within a certain range of consumption, customers with heavy demand but small contribution to the system peak could be excluded from AMI investment/peak shaving investment portfolios, whereas those with a similar demand level but a larger peak contribution can be targeted in such programs as impactful customers. The main contributions of this paper can be summarized as follows:

- A new metric, CMPC, is proposed as a measure for customer segmentation strategy, which accurately assesses the individual customer impact on system peak from a real dataset. We will show that the proposed metric contains different and unique information compared to the existing metrics.
- A three-stage machine learning framework is developed to obtain CMPC for unobservable customers by accurately estimating their contribution to system peak demand.
- The proposed framework is innovative and intuitive, and considers various specific properties of our real data: 1) the linear nature of the relationship between the CMPC and demand level in the same cluster; 2) concentration of residential customers demand within a small range; 3) strong seasonal changes in customer behaviors.
- The proposed framework can handle the uncertainty of the classification process by integrating the probabilistic values for each typical pattern in the regression model.

## II. Data Description and CMPC Definition

### A. Data Description

The available data used in this paper is provided by several mid-west U.S. utilities. The data includes the energy consumption measurements of over 3000 residential customers from SMs, and the corresponding supervisory control and data acquisition (SCADA) data. The data ranges from January 2015 to May 2018 [12]. The SM data was initially processed to eliminate grossly erroneous and missing samples. Accordingly, the data points with a z-score magnitude of larger than 5 are marked as "erroneous" and replaced using local interpolation [13]. The empirical distribution and cumulative distribution function (CDF) of customer monthly energy consumption are obtained and presented in Fig. 1. As shown in the figure, the majority of residential customer monthly consumption samples are concentrated around 1000 kWh, and almost 80% of customers have monthly consumption levels below 1000 kWh. Compared to the industrial and commercial customers, the demand level of residential households is distributed within a smaller range. This indicates that using only demand level for customer segmentation can be a difficult task.

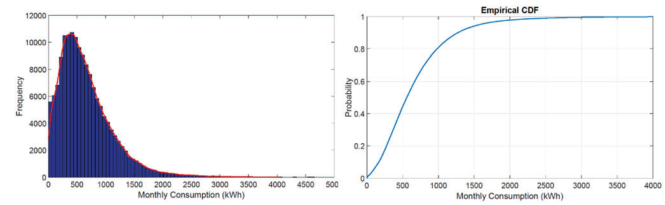



Fig. 1. Monthly consumption distribution: consumption histogram (left), consumption CDF (right).

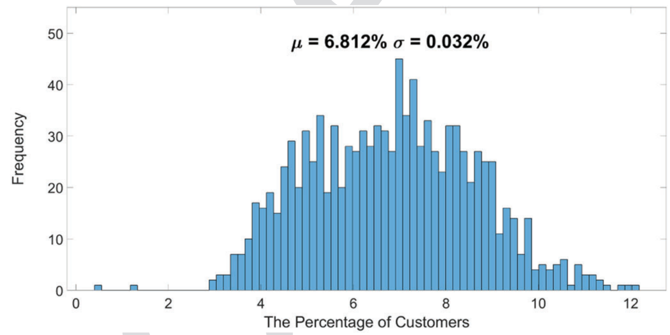



Fig. 2. Percentage of customers whose peak demand coincide with the system peak.

### B. CMPC Definition

The system peak demand is one of the most important operational factors for utilities due to the high marginal cost of energy procurement at the peak time. Hence, it is obligatory to investigate a customer segmentation methodology based on each load's contribution to system peak demand. However, individual customer's peak demand cannot be employed as a measure to assess this contribution, since individual customer peak demand does not necessarily coincide with the system peak. In order to illustrate this, a statistical analysis is performed on the available SM dataset. Fig. 2 shows the percentage of customers whose peak demand coincides with the system peak load. On average only 6% of customers have the same peak time as the system, with a standard deviation of 12%. This means that a customer's peak demand cannot be relied upon to estimate its contribution to the overall system peak load. Thus, in this paper, we propose a new metric, denoted as CMPC, to accurately quantify the contribution of an individual customer to the system peak demand:

$$F_{j,m} = \frac{1}{n} \sum_{d=1}^{n} \frac{p_{j,m}^{d}(t_d)}{P_{m}^{d}(t_d)} \tag{1}$$

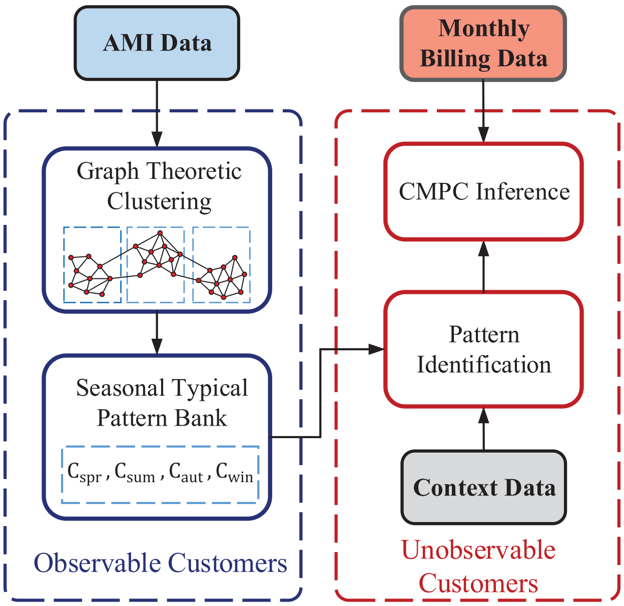
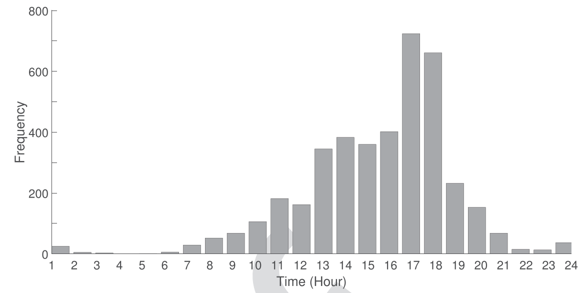
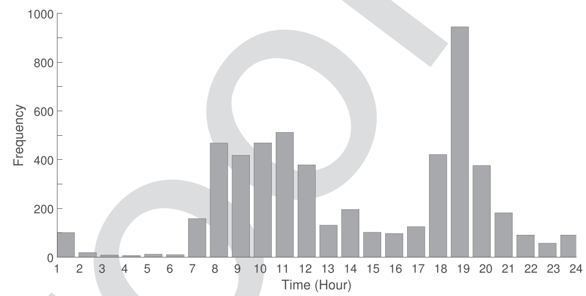



Fig. 3. Proposed data-driven framework.




(a) Summer statistics




(b) Winter statistics

Fig. 4. Seasonal system peak time distribution.

where CMPC of the $j$'th customer at the $m$'th month is denoted by $F_{j,m}$. Here, $p^d_{j,m}(t_d)$ is the customer's demand at time $t_d$ on the $d$'th day of the month, with $n$ denoting the total number of days in the month. Note that $P^d_m$ and $t_d$ are the value and the time of system peak demand on the $d$-th day of the $m$-th month. Hence, CMPC is basically the average customer contribution to the daily system peak demand during a month. A few related but different indices can be found in the literature, such as *coincidence contribution factor*, which is defined as the gap between the aggregate peak demand of a group of customers and their actual consumption at the system peak time [14]. However, the coincidence contribution factor cannot be used as a customer-level metric due to its inability to quantify individual customers' contributions to the system peak load. CMPC can be directly calculated for observable customers using the real-time SM measurements. Considering that not all customers have SMs in practice, especially for residential households, we propose a multi-stage data-driven method for estimating CMPC. The flowchart of the proposed approach is presented in Fig. 3. (I) In the first stage, the demand profiles of observable customers are utilized to build a seasonal consumption pattern bank, $[\{C_{spr}\}, \{C_{sum}\}, \{C_{aut}\}, \{C_{win}\}]$, using a graph theoretic clustering technique. Here, each $\{C_{(\cdot)}\}$ is the set of the typical daily load profiles for a specific season (detailed in Section III). Seasonal data clustering shows a better load behavior identification performance due to its ability to capture the critical seasonal behaviors of customers [15]. (II) Then, a classification module is developed to infer the likelihood of identified seasonal daily consumption profiles for customers without SM data utilizing sociodemographic information. (III) For each typical pattern, a regression model is trained to provide an inference function to estimate the CMPC from customers' monthly billing data. To take into account the variances of CMPC in different typical patterns, a WCR approach is developed based on the results of classification module. Basically, the proposed customer segmentation approach is able to infer CMPC of customers without SMs using their monthly billing information and limited context information.

## III. Graph Theoretical Clustering Algorithm

In this paper, a graph theory-based clustering technique, known as spectral clustering (SC), is adopted. Due to the strong seasonal changes in the customers' behavior, the SC uses seasonal average customer load profiles to identify typical daily load patterns corresponding to different seasons [16], [17]. According to the statistical analysis, both customer behaviors and system peak timing are affected by seasonal changes, as shown in Fig. 4. In Fig. 4(a), the peak time distribution in summer is concentrated around evening interval (17:00–18:00 pm). Meanwhile, the peak time probability rises during daytime and falls sharply at night. One possible reason is the increase of air conditioning usage during summer daytime. In contrast, the peak time distribution of winter is presented in Fig. 4(b). Compared to the summer, the distribution of peak demand time in winter has two concentration points: one in morning hours (8:00–12:00 am), and the other in the evening (18:00–20:00 pm). Also, the peak time probability shows relatively low values during the afternoon interval (13:00–17:00 pm). Hence, in this work, instead of assigning a single pattern to each customer, various patterns are obtained for different seasons to capture the seasonality of customer behaviors [15].

In each season, the AMI dataset is represented as an undirected similarity graph, $G = (V, E)$. $V$ is the set of vertices in the graph, where the $i$'th vertex represents the average daily profile of the $i$'th customer, $V_i = [C^i_1, \ldots, C^i_{24}]$, with $C^i_j$ denoting the average load value at the $j$' hour of day for the $i$'th customer.

$E$ is the set of edges in the graph that connect different vertices, where a non-negative weight, $W_{i,j}$, is assigned to the edge connecting vertices $i$ and $j$. The weight value represents the level of similarity between the two customers' average daily load profiles, with $W_{i,j} = 0$ indicating that the vertices $V_i$ and $V_j$ are not connected. In this paper, the weight $W_{i,j}$ is obtained by adopting a Gaussian kernel function:

$$W_{i,j} = \exp\left(\frac{-||V_i - V_j||^2}{\alpha^2}\right) \tag{2}$$

where $\alpha$ is a scaling parameter that controls how rapidly the weight $W_{i,j}$ falls off with the distance between vertices $V_i$ and $V_j$. To enhance computational efficiency and adaptability to the dataset, we have adopted a localized scaling parameter $\alpha_i$ for each vertex that allows self-tuning of the point-to-point distances based on the local distance of the neighbor of $V_i$ [18]:

$$\alpha_i = ||V_i - V_\varphi|| \tag{3}$$

where, $V_\varphi$ is the $\varphi$'th neighbor of $V_i$, which is selected according to [18]. Therefore, the weight between a pair of points can be re-written as:

$$W_{i,j} = \exp\left(\frac{-||V_i - V_j||^2}{\alpha_i \alpha_j}\right) \tag{4}$$

Given a set of vertices and weight matrix $W = (W_{i,j})_{i,j=1,\ldots,n}$, the clustering process is converted to a graph partitioning problem. In this paper, the objective function of graph partitioning is to maximize both the dissimilarity between the different clusters and the total similarity within each cluster [19]:

$$N(G) = \min_{A_1,\ldots,A_n} \sum_{i=1}^{n} \frac{c(A_i, V \backslash A_i)}{d(A_i)} \tag{5}$$

where, $n$ is the number of vertices, $A_i$ is a cluster of vertices in $V$, $V \backslash A_i$ represents the nodes of set $V$ that are not in set $A_i$, $c(A_i, V \backslash A_i)$ is the sum of the edge weights between vertices in $A_i$ and $V \backslash A_i$, $d(A_i)$ is the sum of the weights of vertices in $A_i$. It has been shown in [16] that the minimum of $N(G)$ is reached at the second smallest eigenvector of the graph's Laplacian matrix, $L$, which can be determined using the weight matrix $W$, as demonstrated in:

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{6}$$

where, $D$ is a diagonal matrix, which $(i, i)$'th element is the sum of $W$'s $i$'th row. The $k$ smallest eigenvalues, $[y_1, y_2, \ldots, y_k]$, of the Laplacian matrix are extracted in the clustering algorithm (see Alg. 1) to build a new matrix $U \in \mathbb{R}^{n \times k}$, where $k$ ranges from 2 to $n$. Leveraging the properties of the graph Laplacians, the data point $V_i$ is reconstructed using the $i$'th row of the $U$ matrix, which enhances the cluster-properties of the data [18]. After data reconstruction, a simple clustering algorithm is able to detect the clusters. In this work, we utilized the $k$-means algorithm to obtain the final solutions from matrix $U$.

Compared to conventional clustering techniques, the SC algorithm has two main advantages: (1) it mainly relies on the weight matrix of the dataset rather than using the high-dimensional demand profile data directly. Also, computing the eigenvalues of matrix $W$ for data reconstruction is equivalent to achieving dimension reduction by employing a linear PCA in a high dimensional kernel space; (2) as a basic idea of SC, graph partitioning problem can be solved without making any assumptions on the data distribution. This improves the robustness of SC, and leads to better clustering performance for complex and unknown data structures [18]. (3) According to equations 2–6, SC converts the clustering process to a graph partitioning optimization problem. Based on *Rayleigh-Ritz theorem*, the solution of this optimization problem is obtained using the $k$ eigenvectors of the Laplacian matrix, which guarantees a good approximation to the *optimal cut*. [20]–[22] The main challenge of SC is that the $k$ value still needs to be determined as a priori. To obtain the optimal $k$, we employ the Davies-Bouldin validation index (DBI), which aims to maximize the internal consistency of each cluster and minimize the overlap of different clusters [23]. The optimal value of $k$ can be obtained when the DBI is minimized. This is shown in Fig. 5 for summer data subset.

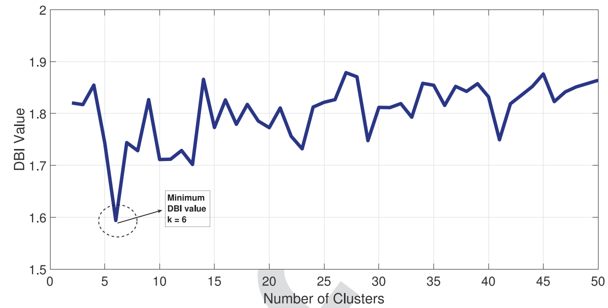



Fig. 5. Cluster validation index performance for summer season.

## IV. CMPC Estimation for Unobservable Customers

In order to assess the CMPC of unobservable customers, a WCR approach is proposed using only their monthly consumption information, as shown in Fig. 6. This framework includes two stages: the first stage is unobservable customer classification based on the seasonal typical consumption pattern bank, and the second stage is cluster-based CMPC inference. It should be noted the two stages cannot be directly combined into one step since they address two different problems.

### A. Unobservable Customer Classification

Since the detailed time-series SM data of unobservable customers is not available, their daily consumption patterns cannot be directly determined beforehand. To link the existing typical load patterns, obtained from the SC technique, to unobservable customers, a pattern classification model is developed. Thus, the goal of this model is to design a classifier that is able to distinguish different behavioral classes based on an input vector that contains sociodemographic information of unobservable customers. The proposed model in this paper maps the sociodemographic information of customers (i.e. working period and dining time) to the typical daily pattern databank. The basic idea is that the typical daily load profiles of customers can be discovered using prior knowledge of their peak consumption timing.

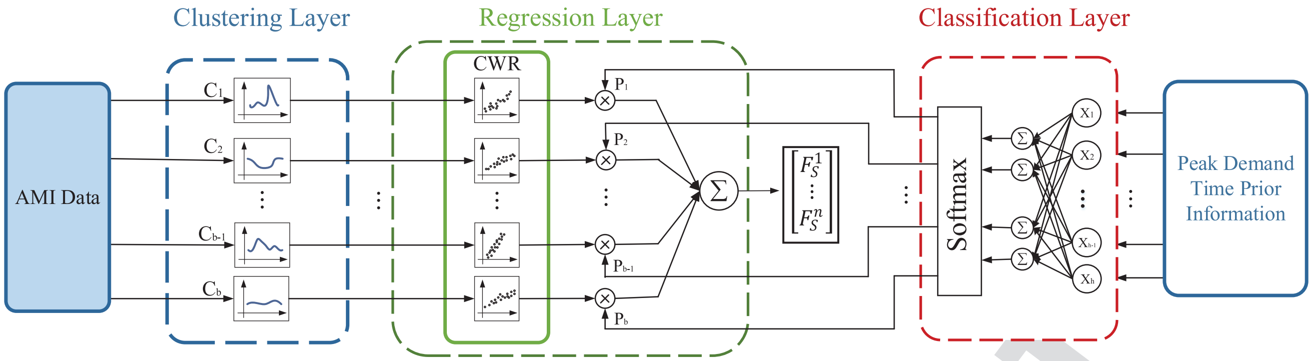



Fig. 6. The structure of WCR model.

Based on the sociodemographic information of customers, the knowledge of customer behavior over a few distinctive intervals in the day can be obtained, namely the morning interval (from 7:00 am to 9:00 am), the afternoon interval (from 12:00 pm to 14:00 pm), and the evening interval (from 18:00 pm to 21:00 pm). This prior information is then used to obtain an approximate probability distribution function of customer peak timing defined as $X^j = \{X_1^j, X_2^j, \ldots, X_{h-1}^j, X_h^j\}$, where $X_i^j$ is the probability of $j$'th customer peak demand occurring at time instant $i$, with $h$ denoting the maximum number of time points. In this work, using the SM measurements of observable customers, $X_i^j$ is determined as follows:

$$X_i^j = \frac{\sum_{d=1}^{n} \Phi(t_d^j)}{n} \tag{7}$$

$$\Phi(t_d^j) = \begin{cases} 1 & \text{for } t_d^j = i \\ 0 & \text{for otherwise} \end{cases} \tag{8}$$

where, $t_d^j$ is the peak demand time of $j$'th customer at the $d$-th day. Thus, the peak timing likelihood distribution, $\{X_1^j, X_2^j, \ldots, X_{h-1}^j, X_h^j\}$, is utilized as the input of the classification model. This classification model for unobservable customers is developed using the multinomial logistic regression (MLR) algorithm. Compared to other binary classification methods such as random forests, MLR is able to obtain the likelihood of different typical profiles for customers rather than picking a single consumption pattern from the databank [23]. The probability that the $j$'th customer follows the $z$'th typical load profile can be written as [24]:

$$P(C_j = z|X^j) = \frac{exp(w_z^T X^j)}{\sum_{j=1}^{k} exp(w_j^T X^j)} \tag{9}$$

where, $C_j$ represents the class of the $j$'th unobservable customer, $T$ is the transposition operator, and $w_z$ is the weight vector corresponding to pattern $z$. The learning parameters $w_z$ are obtained by solving $\nabla_{w_z} J = 0$ over the training set, where $J$ is the classification risk function, defined as follows [25]:

$$J = \sum_{j=1}^{M} \left[ \sum_{z=1}^{k} c_j^z (w_z)^T X^j - \log \sum_{z=1}^{k} \exp((w_z)^T X^j) \right] \tag{10}$$

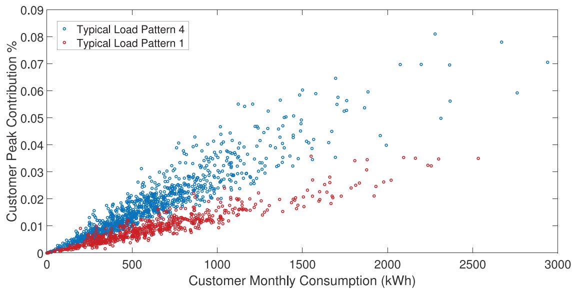
where, $c_j^z$ is the $j$'th element of $c^z$, which is a binary string representing customer class membership. To maximize the log-likelihood function, $J$, with respect to $w_z$, we need to compute the gradient and Hessian of equation (10). Based on the block-structured property of learning parameters and Kronecker product of matrices, the gradient and Hessian of the objective function can be obtained and passed to any gradient-based optimizer to find the maximum a posterior (MAP) estimation of model parameters [26]. In this paper, an iterative reweighted least squares (IRLS) training mechanism was implemented [27]. It should be noted that although there are other methods for performing this maximization, none clearly outperforms IRLS [25].

### B. *Estimation of CMPC for Unobservable Customers*

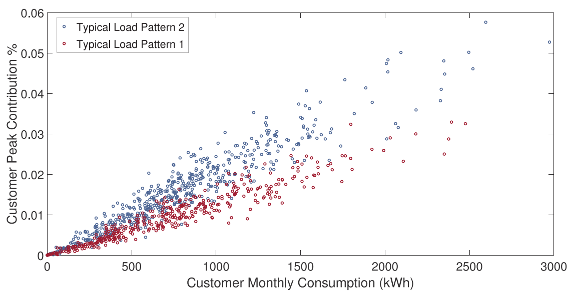
To infer the CMPC for unobservable customers, a WCR model is developed by combining two variables: daily load profile and demand level. The basic idea of WCR approach is to utilize the linear nature of the relationship between the CMPC and monthly energy consumption when the load profiles of customers are similar. This is demonstrated in Fig. 7, where the CMPC and monthly energy consumption of customers in different clusters are shown. As depicted in Fig. 7, the correlation between monthly energy consumption and the CMPC is largely different for customers with two distinct behavioral patterns in the same season.

Hence, for $z$'th typical pattern, a linear regression model is trained for mapping the customer's monthly billing information to the CMPC values. The monthly billing data of consumers is obtained by aggregating their SM data. As shown in Fig. 1 the majority of monthly consumption values are concentrated around 1000 kWh. Then, the actual CMPC value is calculated using the SCADA and SM data at the system peak time. To estimate the parameters $W_z$ and $b_z$ of this regression model, ordinary least square (OLS) is used in this paper [28]. The basic idea is to minimize the sum of the squares of the differences between the estimated and actual CMPCs. The objective function can be written as follows:

$$f_z = min_{W_z, b_z} \sum_{i=1}^{n} (F_{j,m}^i - (E_{j,m}^i W_z + b_z))^2 \tag{11}$$




(a) Monthly energy and CMPC of different patterns in spring




(b) Monthly energy and CMPC of different patterns in summer

Fig. 7. Performance of clusterwise.

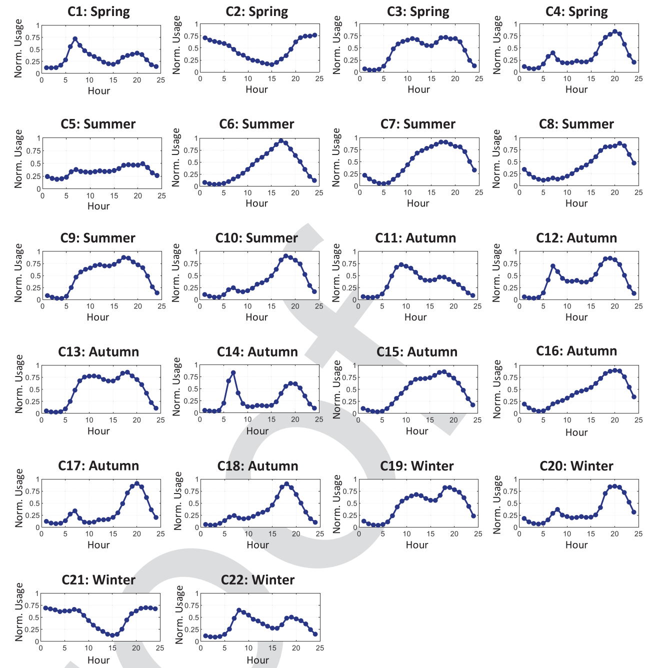



Fig. 8. Seasonal Typical load patterns databank.

where, $E_{j,m}$ and $F_{j,m}$ are the monthly consumption level and the actual CMPC for the $j$'th customer at the $m$'th month. It should be noted that our dataset includes the real SM measurements of over 3000 residential customer and the corresponding SCADA records over 3 years. For each regression model, to reduce the overfitting risk, the dataset is randomly divided into two separate subsets for training (80% of the total data) and testing (20% of the total data). After training, all regression models are then merged into a WCR to estimate the CMPC for unobservable residential customers. Using the cluster probability values obtained from the classification model, $P(C_j = z|X^j)$, the estimated CMPC for the $j$'th customer at the $m$'th month, $\hat{F}_{j,m}$, is determined as follows:

$$\hat{F}_{j,m} = \sum_{z=1}^{k} P(C_j = z|X^j)(W_z E_{j,m} + b_z) \tag{12}$$

Hence, the proposed WCR is able to estimate the CMPC of unobservable customers using only their measured monthly consumption within a probabilistic classification setting. OLS regression can produce unbiased estimates that have the smallest variance among all possible linear estimators if the model follows several basic assumptions to satisfy the conditions of Gauss-Markov theorem [29]. In our work, the linear nature of the relationship between the CMPC and monthly energy consumption in the same cluster and random selection of training data help satisfy these assumptions, thus ensuring the theoretical performance of WCR. Also, it should be noted that in general the performance of the OLS is impacted by outliers and extreme observations [28]. However, in our problem outliers and extreme values are highly unlikely since the residential customers' monthly demand levels are concentrated within a small range; almost 80% of customers have monthly consumption levels below 1000 kWh.

## V. Numerical Results

The real distribution system provided by our utility collaborator is equipped with SMs, thus fully observable. This enables us to calculate the exact CMPC of each customer. To test the proposed customer segmentation method for partially observable systems, we assume that 20% of customers are unobservable and then compare the estimation results with the actual CMPCs. Thus, the data of observable customers (the remaining 80% of the total data) is divided into 4 subsets corresponding to different seasons of the year for model training.

### A. SC Algorithm Performance

For every subset, the optimal cluster number is determined using DBI and typical load patterns are obtained employing the SC algorithm (detailed in Section III). Fig. 8 and Fig. 9 present the 22 typical load shapes, namely $C_1$, $C_2$, ..., $C_{22}$, and the distribution of population of customers belonging to each cluster during all the seasons. As shown in the figures, the number of typical load profiles in different seasons is not the same and the SC approach is able to capture the critical seasonal consumption patterns. In spring, around 22% of customers show typically higher consumption levels during the morning (around 7:00 am). In contrast, more than 38% of customers have higher energy consumption during the evening (around 20:00 pm). Meanwhile, more than half of customers present low energy consumption value during the afternoon period. The typical load profiles in summer are different from spring. Except for $C_5$, the typical load patterns of 85% of all customers show similar behavioral tendencies. This could be due to air-conditioning

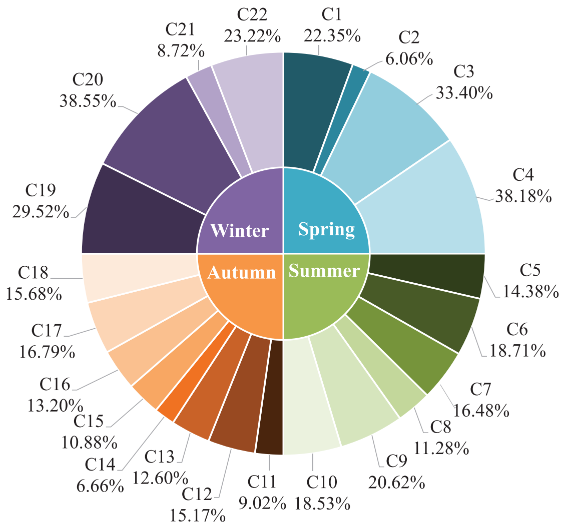



Fig. 9. Proportion of typical load patterns for different seasons.

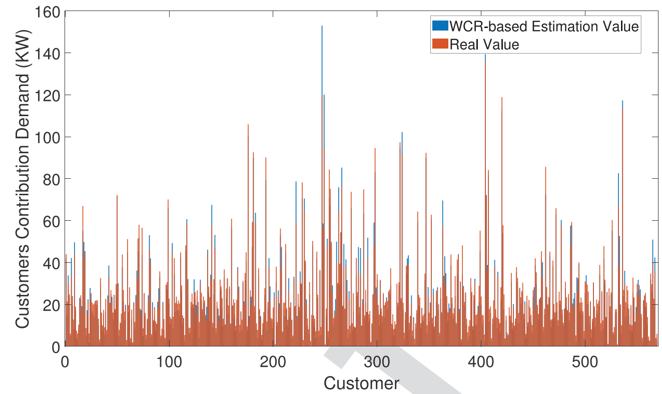



Fig. 10. Comparison of WCR-based estimation value and real value.

load consumption during time intervals with higher temperature. Based on the typical load patterns, the majority of peak demand occurs during the evening interval. For around 74% of customers in summer, the peak time ranges from 17:00 pm to 19:00 pm. In fall, the number of typical load patterns is relatively larger rather than other seasons due to variability of customer behavior. Compared to summer, when peak demand barely happens in the morning, more than 40% of customers have high consumption at around 7:00 am in fall, such as $C_{11}, C_{12}, C_{13}$ and $C_{14}$. Also, around 23% of customers provide almost zero consumption from 10:00 am to 15:00 pm, and nearly one-third of customers show two peaks in the morning and evening periods. The winter typical daily patterns are similar to the results of spring since these two seasons have similar weather in mid-west U.S.

### *B. WCR Performance*

When the seasonal consumption pattern bank is developed using the SM data of observable customers, the WCR models are utilized to infer the CMPC of unobservable customers.

*1) Classification Performance Analysis:* For the classification part, the Area under the Curve (AUC) index is employed to assess the performance of MLR model [30]. AUC is determined as follows:

$$\gamma = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \tag{13}$$

where, TP is the True Positive, TN is the True Negative, FP is the False Positive, FN is the False Negative, and N is the number of total Negatives. Compared to the commonly-used metric, accuracy, the AUC does not depend on the cut-off value that is applied to the posterior probabilities to evaluate the performance of a classification model [31].

The meaningful range of AUC is between 0.5 to 1. In order to avoid the overfitting problem, the $k$-fold cross-validation method is applied to the MLR to ensure the randomness of the training set [32]. Based on the prior information on customer peak timing distribution, the MLR achieves an AUC value of 0.7 when assigning daily load patterns to unobservable customers.

*2) Regression Performance Analysis:* Based on the WCR approach, the CMPC of unobservable customers can be estimated using the monthly billing data. Fig. 10 shows the performance of WCR by comparing the actual CMPC with the estimated CMPC for each customer in the testing set for one month. As can be seen, the estimated values are able to accurately track the unobservable customer's real contribution to system peak demand. To assess the performance of the model, the goodness-of-fit measure, $R^2$, and the mean absolute percentage error (MAPE) are utilized in this paper. These two indices are presented in Table I for all seasons. Based on these results, the regression model has a good performance for estimation of CMPC of unobservable customers in this case.

TABLE I
PERFORMANCE OF SEASONAL WCR MODELS WITH $R^2$ AND MAPE

| Season | Average $R^2$ | Average MAPE |
|---|---|---|
| Spring | 0.9446 | 12.44% |
| Summer | 0.9071 | 14.24% |
| Fall | 0.9384 | 13.18% |
| Winter | 0.9204 | 13.7% |

### *C. Metric and Method Comparison*

In this section, we demonstrate that the proposed segmentation strategy can target suitable customers, which cannot be classified by existing method in the literature, including customer peak demand-based and load profile entropy-based segmentation strategies [6], [8]. Furthermore, to validate the performance of our multi-stage machine learning framework, we have compared the peak contribution estimation MAPE of the proposed learning-based framework with previous method [33].

*1) Comparing Customer Peak Demand-Based Strategy and Proposed Method:* Customer peak demand is a conventional index to describe the potential impact of individual customers on the overall peak demand, which is commonly-used by utilities to perform customer segmentation [8]. In Fig. 11, the difference between the proposed CMPC and customer peak demand values are presented. It can be seen that the customer peak demand values are generally much higher than CMPC values due to the

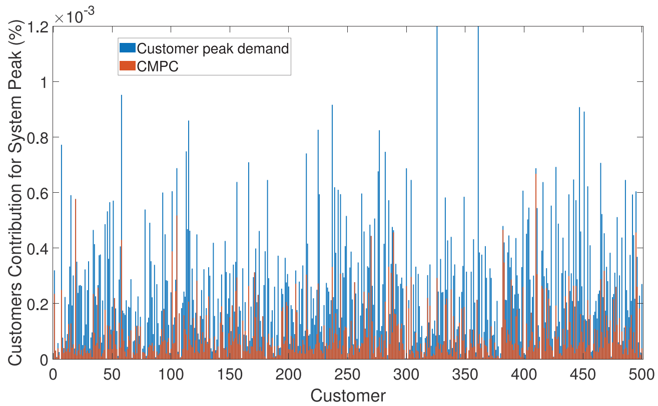



Fig. 11. Comparison of CMPC and customer peak demand.

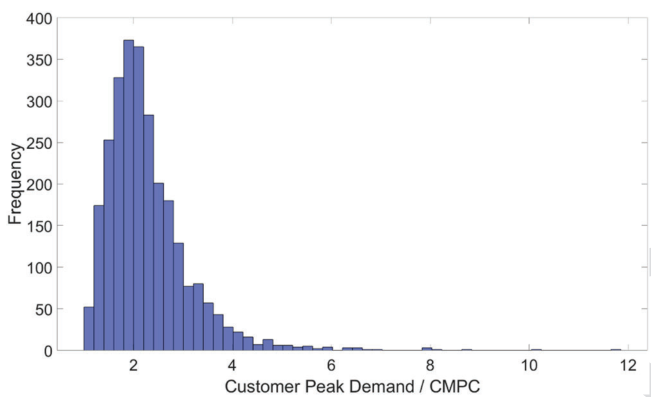



Fig. 12. The histogram of customer peak demand over CMPC ratio.

diversity of load behaviors. According to Fig. 12, the customer's peak demand can reach five times the customer's actual contribution to the system peak. This considerable difference shows that compared to the proposed method, customer peak demand-based strategy is a very conservative method of quantifying the actual impact of customers, which could lead to unnecessary over-investments in AMI expansion.

*2) Comparing Load Profile Entropy-Based Strategy and Proposed Method:* Entropy is a measure of the variability and uncertainty of customer demand, which has been used to develop customer segmentation approach for peak shaving program targeting [6]. Customers with lower entropy levels have stable consumption behaviors, which makes them higher priority candidates for peak reduction. In Fig. 13, the relationship between CMPC and entropy is presented. It is observable that customers with high CMPC do not necessarily have low entropy values. This indicates that these two concepts are almost uncorrelated and do not contain mutual information. Hence, unlike the proposed method, the entropy-based strategy does not provide information about customers' impact on system peak demand, and thus, cannot be used as a generic strategy for guiding peak shaving/AMI planning.

*3) Comparing the Performance of the Proposed Multi-Stage Machine Learning-Based Framework With an Existing Method:* The performance of the proposed multi-stage machine learning framework is compared with an existing baseline method [33] in terms of estimation accuracy. The baseline method uses ordinary

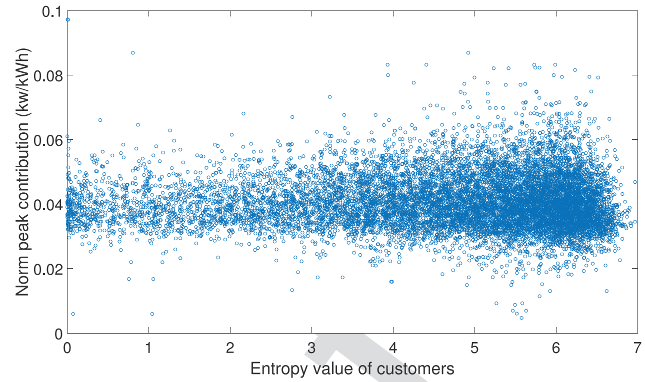



Fig. 13. The relationship between CMPC and entropy.

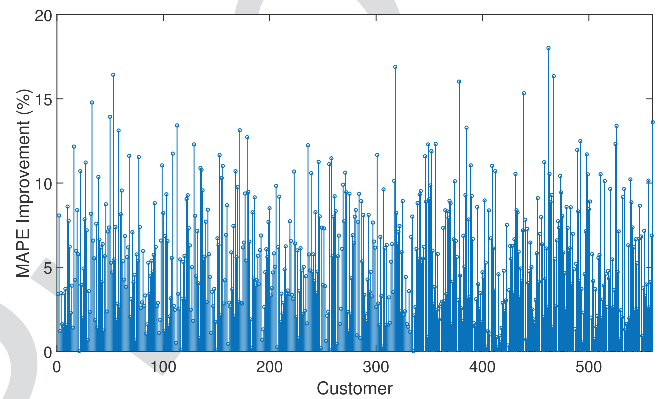



Fig. 14. Comparison of proposed method and existing method [33].

least square regression to determine the peak demand based on the periodic energy consumption. As shown in Fig. 14, the estimation MAPE values for our proposed method are generally lower than the results obtained from the previous method in [33]. Our framework has been able to improve the estimation MAPE by 5% on average. Furthermore, a maximum point-wise improvement level of 18% has been achieved over the previous baseline method. Hence, based on this AMI dataset, the proposed method shows a better estimation accuracy compared to the previous work.

### D. Application of the Proposed Metric and Strategy

To evaluate the performance of the proposed metric and the strategy in system operation, we have applied our works to a basic direct load control-based demand response model, which gives utilities the option to remotely shut down appliances during daily peak demand periods [34]. A 300-house radial distribution network has been considered to evaluate the performance of different segmentation strategies. 35% of unobservable customers are selected for meter installation and participation in peak shaving using three different segmentation metrics: 1) select residential candidates randomly (base strategy); 2) select residential candidates by ranking monthly demand level; 3) select residential candidates based on the proposed CMPC. According to the existing works [35], [36], we have assumed average load elasticity of customers to be 0.21 p.u. We have the compared daily peak reductions in one month (28 days) under the three different customer segmentation strategies. As

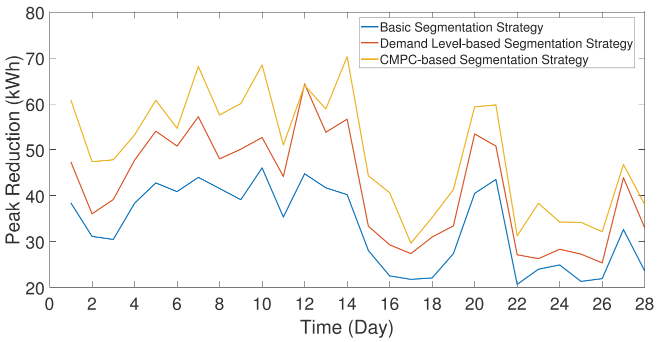



Fig. 15. Comparison of peak reduction using three different segmentation strategies.

shown in Fig. 15, using the proposed CMPC strategy, over 1400 kWh peak demand has been saved in this month, which is higher than the other two segmentation strategies. Specifically, in this case, when basic and demand level-based strategies are replaced by CMPC-based strategy, the average peak reduction increases by 50.4% and 19.7%, respectively. Thus, by comparison, the proposed customer segmentation strategy and the CMPC metric have the potential to provide enhanced customer targeting guidelines for improving operational frameworks. As a future research direction, we will utilize the proposed metric in more advanced and detailed operation models.

## VI. Conclusion

In this paper, we have presented a new metric for customer segmentation, CMPC, which can quantify the contributions of individual customers to system peak demand. Moreover, to accurately estimate the CMPC of unobservable residential customers, an innovative three-stage machine learning framework is developed using only their monthly billing data. Employing our real SM data, it is demonstrated and validated that the proposed metric provides utilities with additional actionable information for customer segmentation compared to the existing metrics. This segmentation strategy helps utilities effectively identify impactful customers from thousands of unobservable customers for investment decisions, such as AMI expansion. Also, these customers can be targeted as candidates for residential-level demand-side management (DSM) programs to reduce the critical system peak demand, thus, decreasing the high marginal cost and the risk of system failure. Our work offers other potential benefits for utilities. For example, recently, utilities have been showing increasing interest in residential-level retail price design due to the significant contribution of residential customers to the system peak. The proposed CMPC, together with the developed machine learning framework, can provide a reasonable strategy to obtain guidelines for retail price design by accurately quantifying the impact of residential customers on the system.

## References


[1] R. Gulbinas, A. Khosrowpour, and J. Taylor, "Segmentation and classification of commercial building occupants by energy-use efficiency and predictability," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1414–1424, May 2015.

[2] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "A multi-timescale data-driven approach to enhance distribution system observability," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 3168–3177, Jul. 2019.

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

[4] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen, "Clustering of connection points and load modeling in distribution systems," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1255–1265, May 2013.

[5] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016.

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

[7] A. Kavousian, R. Rajagopal, and M. Fischer, "Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings," *Energy Buildings*, vol. 99, pp. 220–230, Apr. 2015.

[8] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 420–430, Jan. 2014.

[9] Y. Yu *et al.*, "Good consumer or bad consumer: Economic information revealed from demand profiles," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2347–2358, May 2018.

[10] M. Sun, Y. Wang, G. Strbac, and C. Kang, "Probabilistic peak load estimation in smart cities using smart meter data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1608–1618, Feb. 2019.

[11] M. Uddin, M. F. Romlie, M. F. Abdullah, S. A. Halim, A. H. A. Bakar, and T. C. Kwang, "A review on peak load shaving strategies," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 3323–3332, 2018.

[12] Z. Wang, "Dr. Zhaoyu Wang's home page," [Online]. Available: http://wzy.ece.iastate.edu/Testsystem.html Q2

[13] D. Cousineau and S. Chartier, "Outlier detection and treatment: A review," *Int. J. Psychol. Res.*, vol. 3, no. 1, pp. 58–67, Jan. 2010.

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

[15] K. Chen, J. Hu, and Z. He, "Data-driven residential customer aggregation based on seasonal behavioral patterns," in *Proc. IEEE Power Energy Soc. General Meeting*, Jul. 2017, pp. 1–5.

[16] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural Inf. Process. Syst.*, pp. 849–856, 2002. Q3

[17] K. Dehghanpour, Y. Yuan, Z. Wang, and F. Bu, "A game-theoretic data-driven approach for pseudo-measurement generation in distribution system state estimation," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 5942–5951, Nov. 2019.

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

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

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

[21] D. Spielman and S. Teng, "Spectral partitioning works: Planar graphs and finite element meshes," in *Proc. 37th Annu. Symp. Found. Comput. Sci.*, 1996, pp. 1–1.

[22] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. London, U.K.: Chapman & Hall, 2013.

[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] Z. Xu, Z. Hong, Y. Zhang, J. Wu, A. C. Tsoi, and D. Tao, "Multinomial latent logistic regression for image understanding," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 973–987, Feb. 2016.

[25] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[26] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[27] T. Minka, "A comparison of numerical optimizers for logistic regression," Dept. of Statistics, Carnegie Mellon Univ., Tech. Rep., pp. 1–1, 2003. Q4

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

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Berlin, Germany: Springer, 2013.
[30] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[31] D. Thorleuchter and D. V. den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 13 026–13 034, Dec. 2012.
[32] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[33] W. H. Kersting, *Distribution System Modeling and Analysis*. Boca Raton, FL, USA: CRC Press, 2016.
[34] C. Chen, J. Wang, and S. Kishore, "A distributed direct load control approach for large-scale residential demand response," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2219–2228, Sep. 2014.
[35] S. Gyamfi, S. Krumdieck, and T. Urmee, "Residential peak electricity demand responsehighlights of some behavioural issues," *Renew. Sustain. Energy Rev.*, vol. 25, pp. 71–77, 2013.
[36] C. Vivekananthan, Y. Mishra, G. Ledwich, and F. Li, "Demand response for residential appliances via customer reward scheme," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 809–820, Mar. 2014.

**Yuxuan Yuan** (Student Member, IEEE) received the B.S. degree in electrical and computer engineering in 2017 from Iowa State University, Ames, IA, USA, 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** (Member, IEEE) received the B.Sc. and M.S. degrees from the University of Tehran in electrical and computer engineering, 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. His research interests include machine learning and data mining for monitoring and control of smart grids, and market-driven management of distributed energy resources.

**Fankun Bu** (Student Member, IEEE) 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 was a Commissioning Engineer with NARI Technology Co., Ltd., Nanjing, China. From 2013 to 2017, he was an Electrical Engineer with State Grid Corporation of China, Nanjing, China. His research interests include distribution system modeling, smart meter data analytics, renewable energy integration, and power system relaying.

**Zhaoyu Wang** (Member, IEEE) 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 Georgia Institute of Technology, Atlanta, GA, USA, in 2012 and 2015, respectively. He is currently the Harpole-Pentair Assistant Professor with Iowa State University, Ames, IA, USA. His research interests include power distribution systems and microgrids, particularly on their data analytics, and optimization. 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 is the Secretary of the IEEE Power and Energy Society (PES) Award Subcommittee, Co-Vice Chair of PES Distribution System Operation and Planning Subcommittee, and Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Editor for the IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE PES LETTERS and IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, and an Associate Editor for *IET Smart Grid*.