A Data-Driven Framework for Assessing Cold Load Pick-Up Demand in Service Restoration

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

***Abstract*—Cold load pick-up (CLPU) has been a critical concern to utilities. Researchers and industry practitioners have underlined the impact of CLPU on distribution system design and service restoration. The recent large-scale deployment of smart meters has provided the industry with a huge amount of data that are highly granular, both temporally and spatially. In this paper, a data-driven framework is proposed for assessing CLPU demand of residential customers using smart meter data. The proposed framework consists of two interconnected layers: 1) At the feeder level, a nonlinear autoregression model is applied to estimate the diversified demand during the system restoration and calculate the CLPU demand ratio. 2) At the customer level, Gaussian mixture models and probabilistic reasoning are used to quantify the CLPU demand increase. The proposed methodology has been verified using real smart meter data and outage cases.**

***Index Terms*—Cold load pick-up, distribution systems, service restoration, least squares support vector machine, Gaussian Mixture Model.**

I. Introduction

COLD load pick-up (CLPU) is a challenging issue in electric power industry [1]–[3]. The CLPU demand is an increased load at service restoration phase due to the loss of load diversity. During normal system operation, the on-off switching cycles of thermostatically controlled loads (TCL) within a population of customers take place independently because of the heterogeneity of appliances and the diversity of customer behaviors. However, immediately after a long power outage in the restoration phase, the switching cycles of TCLs will coincide and become highly correlated for a period of time. This phenomenon is the main reason for the abnormal level of demand due to the temporary lack of diversity. In this paper, the term "CLPU demand" refers to this undiversified load during the restoration.

For feeders with high penetration of TCLs, CLPU can have serious consequences, such as restoration failure [4]–[8], transformers aging [9], [10], transformer overloading [11], and unacceptable voltage drops [12]. CLPU demand can continue several minutes to even several hours after extreme weather conditions [2]. Hence, it is necessary to quantify the impact of CLPU on distribution system design and restoration. To achieve this goal, the primary task is to quantify the deviation of CLPU demand from normal (diversified) load in historical outage cases. This will help the utilities to extract useful information for future service restorations.

Previous papers have mainly focused on model-driven methods for CLPU demand assessment. In [1], a physical model was built for simulating steady and transient response of thermostatically-controlled residential electric space heating devices. Based on the developed model, the aggregate impacts of space heaters on feeder-level CLPU demand were analyzed. In [13], a simple and practical model was developed to represent temperature dynamics in a house with a thermostatically-controlled heater/air-conditioner. The model can be used in load management and aggregate CLPU impact evaluation. In [14], similar groups of elementary component load models were built and the load models in the same group were aggregated by using statistical techniques to simulate CLPU. In [15], a multi-state physical load model was developed to capture the behavior of end-use loads. Besides using air temperature as a control signal, other variables, such as price, can also be integrated into the model. In [3], the developed model in [15] was used to account for the multi-state operation of residential heat pumps. This model was then employed to estimate the magnitude and duration of CLPU demand. In [16], CLPU demands of seven houses with different types of electric heating equipment were measured, and field studies were also performed for load restoration process. Although [16] used field measurements to analyze CLPU demand, the employed dataset was procured from the measurements of only a limited number of residential customers, and it fails to employ estimation methods to capture the variations in the expected diversified demand at the time of restoration. Thus, previous works are largely dependent on detailed dynamic modeling of residential/commercial appliances.

While model-driven methods for CLPU demand evaluation offer benefits, such as physical interpretability and cost-efficiency, their disadvantages cannot be ignored. Residential loads depend on many factors, such as the types of appliances, the states of appliances, customer behaviors, and house thermal resistance and capacitance. Therefore, to accurately model a house load, enough detailed information should be collected, which is very challenging to accomplish for utilities in practice. This lack of such detailed information can lead to considerable

Manuscript received October 10, 2018; revised January 31, 2019 and April 18, 2019; accepted May 25, 2019. This work was supported by the Advanced Grid Modeling Program at the U.S. Department of Energy Office of Electricity under DE-OE0000875. Paper no. TPWRS-01553-2018. *(Corresponding author: Zhaoyu Wang.)*

Q1

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: fbu@iastate.edu; kavehdeh1@gmail.com; wzy@iastate.edu; yuanyx@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.2019.2922333

modeling bias. On the other hand, in the past decade, smart meter data with high temporal-spatial granularity has become widely available to utilities [17], which provides an opportunity to address the shortcomings of previous model-based approaches. Specifically, the impact of all the aforementioned unknown factors can be reflected in the smart meter data. For example, a larger thermal inertia value results in a decrease in the rate of change of indoor temperature that leads to less power consumption at the time of restoration, which is then measured by the smart meters. Hence, utilities can obtain CLPU information from smart meter data with high fidelity, instead of relying on model-driven techniques, which need detailed explicit knowledge of thermal parameters and appliances' information. Considering this, in this paper, we will develop a data-driven framework to assess feeder-level CLPU demand ratio and customer-level CLPU demand increase. Compared with the model-driven methods, the framework can provide an approach for assessing CLPU demand without the need for developing specific load models.

The main contribution of this work is to propose a framework for assessing CLPU demand at the time of restoration, which is used for developing statistical CLPU models. The proposed framework consists of two layers: 1) At the feeder level, the ratio of the undiversified CLPU demand to the estimated diversified demand is obtained. To achieve this, a least squares support vector machine (LS-SVM) auto-regression model is employed to estimate the diversified demand under the assumption that the outage has not happened. Then the CLPU demand ratio is calculated by dividing the actual undiversified CLPU demand at the time of restoration by the estimated diversified demand. Finally, a CLPU ratio regression model is developed based on the obtained historical CLPU ratios under different outage duration and ambient temperatures. The developed regression model can be used for predicting how the load behaves under new and previously unseen outage cases. Therefore, one innovative aspect of this paper is using a load estimation technique to assess feeder-level CLPU demand, which has not been applied in previous papers regarding CLPU demand evaluation. 2) At the customer level, a novel CLPU demand increase assessment approach is proposed. Gaussian mixture models (GMM) are applied to devise a probabilistic technique towards constructing marginal probability density functions (PDF) of customer demand increase due to CLPU. Then the customer CLPU demand increase is analyzed statistically for a set of customers to evaluate the loss of load diversity. The performance of the developed framework is verified using real smart meter data from three Midwest U.S. utilities. We have also shown that using the proposed approach the PDF of demand increase due to CLPU can be estimated for any group of customers, which can provide an invaluable guideline for designing sequential restoration plans for distribution systems.

The rest of the paper is organized as follows: Section II introduces the proposed framework of CLPU demand assessment and the real dataset. Section III presents the procedure of feeder CLPU demand ratio evaluation. In Section IV, the procedure of assessing customer CLPU demand increase is presented. In Section V, outage case studies are analyzed. Section VI concludes the paper.

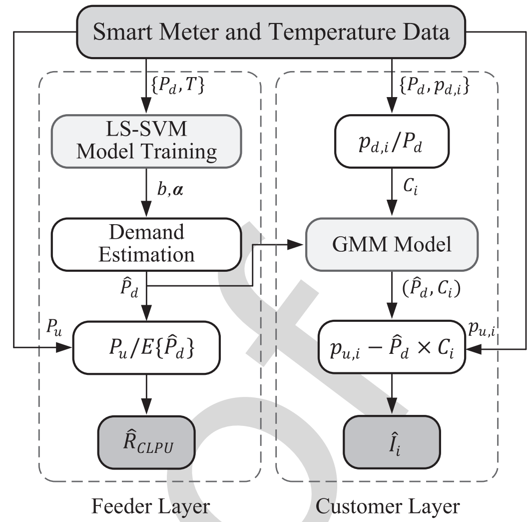


Fig. 1. CLPU demand assessment framework.

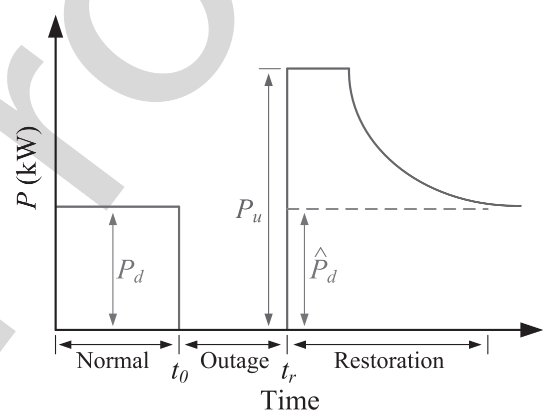


Fig. 2. Typical CLPU demand curve.

II. Proposed CLPU Demand Assessment Framework and Real Dataset Description

This section presents a high-level overview of the proposed framework for feeder-level and customer-level CLPU demand assessment. We will also describe the real smart meter dataset available to us. Assuming customers' smart meter data during normal and restoration conditions is available to utilities, then the feeder-level demand can be obtained by aggregating time-aligned customer-level demand. The overall framework is shown in Fig. 1.

A. Feeder Layer

The objective of feeder-level CLPU demand assessment is to obtain the CLPU demand ratio, R_{CLPU}, defined as follows:

$$R_{CLPU} = \frac{P_u}{P_d} \tag{1}$$

where, P_u is the undiversified feeder demand at the time of restoration, t_r, after outage occurrence at t_0, and P_d is the diversified feeder demand at time t_r. However, as shown in Fig. 2, the actual measured feeder demand at the time of restoration is undiversified due to CLPU. This implies that P_d at time t_r cannot be directly obtained from smart meter data, and needs to be estimated based on the demand data history during normal system operation. Hence, the first task at this level is to estimate P_d at the time of restoration. For this, we design a machine learning technique which is based on a nonlinear auto-regression model with exogenous input (NARX) and is implemented using LS-SVM. The LS-SVM is trained using historic P_d and temperature data to continuously predict the future diversified demand. Since the actual feeder load at the time of restoration is undiversified due to CLPU, the outcome of the machine learning model is the estimated diversified feeder demand *if the outage did not happen*. Therefore, taking the estimation residuals into account, the estimated diversified feeder demand $\hat{P}_d$ at time t_r is modeled as follows:

$$\hat{P}_d = P_d + \varepsilon_{t_r} \tag{2}$$

where, ε_{t_r} denotes the machine learning estimation residual. Assuming that the residual follows a Gaussian distribution, namely $\varepsilon_{t_r} \sim \mathcal{N}(0, \sigma^2)$, with σ defining the machine learning framework uncertainty, the estimated diversified demand also follows a Gaussian distribution [18]:

$$\hat{P}_d \sim \mathcal{N}(P_d, \sigma^2) \tag{3}$$

Therefore, the CLPU demand ratio can be estimated as follows:

$$\hat{R}_{CLPU} = \frac{P_u}{E\{\hat{P}_d\}} \tag{4}$$

where, $E\{\cdot\}$ is the empirical averaging operator.

B. Customer Layer

The objective of this level is to construct the marginal PDF of individual customer demand increase due to CLPU at the time of restoration, denoted as I_i for the i^{th} customer and defined as follows:

$$I_i = p_{u,i}(t_r) - p_{d,i}(t_r) \tag{5}$$

where, $p_{u,i}(t_r)$ is the actual customer demand corresponding to the undiversified feeder load, and $p_{d,i}(t_r)$ is the customer demand at the time of restoration *if the outage did not happen*. Similar to the variable P_d in equation (1), since $p_{d,i}(t_r)$ is unknown and cannot be measured by smart meters, it needs to be estimated. However, compared to feeder demand, customer demand can be much more volatile. Considering this volatility, instead of directly estimating $p_{d,i}(t_r)$ for each customer, this paper adopts a probabilistic learning approach to construct the marginal PDF of the estimated customer demand at time t_r using the obtained $\hat{P}_d$ for the feeder. Based on the demand data history from smart meters, a *contribution factor* is defined for the i^{th} customer, denoted as C_i, which determines the customer contribution to feeder demand (P_d). Note that C_i is obtained in normal state (without outage) when $p_{d,i}(t)$ can be measured directly by smart meters, as follows:

$$C_i(t) = \frac{p_{d,i}(t)}{\sum_{j=1}^{M} p_{d,j}(t)} = \frac{p_{d,i}(t)}{P_d(t)} \quad i = 1, \ldots, M. \tag{6}$$

where, M is the total number of customers connected to the feeder. Hence, during the normal operation, an individual customer demand can be determined as $p_{d,i}(t) = P_d(t)C_i(t)$. Noting the dependency of $p_{d,i}(t)$ on both $P_d(t)$ and $C_i(t)$, to obtain the marginal PDF of $\hat{p}_{d,i}(t_r)$, the joint PDF of the estimated diversified feeder demand ($\hat{P}_d(t)$) and the contribution factor at the time of restoration is constructed. This joint PDF is determined using a GMM technique, which employs past customer demand measurements and the corresponding estimated diversified feeder demand. It will be shown that a nonlinear transformation of this joint PDF can be used to obtain the marginal PDF for $\hat{p}_{d,i}(t_r)$. The CLPU demand increase for the i^{th} customer at the time of restoration, $\hat{I}_i$, is estimated as:

$$\hat{I}_i = p_{u,i}(t_r) - \hat{p}_{d,i}(t_r) \tag{7}$$

Note that given the obtained marginal PDF for $\hat{p}_{d,i}$, equation (7) also leads to a marginal PDF for the individual customer demand increase.

C. Real Dataset Description

The available smart meter data history contains three U.S. Midwest utilities' energy consumption data (kWh) of over 10,000 residential customers with a 15-minute time resolution, and the time range is about four years. This data includes time stamps which have been used for customer-level demands' time-alignment. When an outage occurs, the meter will keep a record of outage start time, end time, and the associated energy consumption readings. The ambient temperature data is obtained from the National Oceanic and Atmospheric Administration (NOAA) database [19] , and is time-aligned with smart meter data.

III. Feeder-Level Diversified Demand Estimation

In this section, a LS-SVM regression model is developed to estimate the diversified feeder demand, $\hat{P}_d$, at the time of restoration, which is then used to determine the CLPU demand ratio, R_{CLPU}.

The LS-SVM is based on a support vector margin maximization process to minimize the machine learning structural risk function. This regression model has many advantages, including good generalization capability and low susceptibility to local minima [18], [20], and has been employed in distribution systems [21], [22]. In demand estimation, the selection of explanatory variables is critical. The feeder demand at a certain time is primarily affected by the temperature at that time and is highly correlated with previous demand samples within a certain time period [18]. Demand also changes with seasons and days of week (working day vs non-working day). To capture seasonal and daily demand diversity, the dataset is divided across seasons and working/non-working days, respectively. The explanatory variable $\mathbf{x}(t) \in \mathbb{R}^n$, which acts as the input to the demand estimation model, is built as follows:

$$\mathbf{x}(t) = [P_d(t-1), \ldots, P_d(t-n_{lag}), T(t)]^\mathsf{T} \tag{8}$$

where, $P_d(t-i)$ is the feeder demand at time $t-i$, n_{lag} is the maximum time lag, and $T(t)$ is the ambient temperature at time t. Therefore, using this explanatory variable, feeder demand at time t can be expressed as:

$$P_d(t) = \boldsymbol{\omega}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t)) + b + \varepsilon_t \tag{9}$$

where, $\boldsymbol{\omega} \in \mathbb{R}^{n_h}$ and $b \in \mathbb{R}$ represent regression model parameters, $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ is a mapping function, transforming low dimensional input vector $\mathbf{x}(t)$ into a high dimensional vector $\boldsymbol{\varphi}(\mathbf{x}(t))$, and ε_t is a normally distributed random variable representing the estimation residual.

Given the current feeder demand and temperature values, together with the past feeder demand samples with certain time lags, a training set of size N_{tr} can be developed, $\mathbf{D}_{tr} = \{\mathbf{x}(t), \mathbf{P}_d(t)\}_{t=1}^{N_{tr}}$. To obtain the optimal values of learning parameters, a structural risk function, J, is formulated and minimized with respect to $\boldsymbol{\omega}$, b, and ε_t, over the training set. This optimization is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\omega}, b, \varepsilon_t} J = \frac{1}{2}\boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega} + \gamma \sum_{t=1}^{N_{tr}} \varepsilon_t^2 \\ &\text{s.t. } P_d(t) = \boldsymbol{\omega}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t)) + b + \varepsilon_t, \quad t = 1, \ldots, N_{tr}. \end{aligned} \tag{10}$$

where, γ is a regularization constant to prevent overfitting. To solve this optimization problem, the Lagrangian, $\mathcal{L}$, is constructed as a function of regression parameters:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\omega}, b, \varepsilon_t; \boldsymbol{\alpha}) = J(\boldsymbol{\omega}, b, \varepsilon_t) - \\ \sum_{t=1}^{N_{tr}} \alpha_t(\boldsymbol{\omega}^T\boldsymbol{\varphi}(\mathbf{x}_t) + b + \varepsilon_t - P_d(t)) \end{aligned} \tag{11}$$

where, α_t's are Lagrange multipliers. The optimality conditions are obtained by solving $\nabla_{(\boldsymbol{\omega}, b, \varepsilon_t)}\mathcal{L} = 0$, as follows:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}} = 0 \rightarrow \boldsymbol{\omega} = \sum_{t'=1}^{N_{tr}} \alpha_{t'}\boldsymbol{\varphi}(\mathbf{x}(t')) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{t=1}^{N_{tr}} \alpha_t = 0 \\ \frac{\partial \mathcal{L}}{\partial \varepsilon_t} = 0 \rightarrow \alpha_t = \gamma\varepsilon_t \\ \frac{\partial \mathcal{L}}{\partial \alpha_t} = 0 \rightarrow P_d(t) = \boldsymbol{\omega}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t)) + b + \varepsilon_t \end{cases} \tag{12}$$

Combining equations (9) and (12), $\boldsymbol{\omega}$ can be eliminated from the regression model as shown below:

$$P_d(t) = \sum_{t'=1}^{N_{tr}} \alpha_t \boldsymbol{\varphi}(\mathbf{x}(t'))^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t)) + b + \frac{\alpha_t}{\gamma} \quad t = 1, \ldots, N_{tr}. \tag{13}$$

The term $\boldsymbol{\varphi}(\mathbf{x}(t'))^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t))$ in equation (13) can be represented by a *kernel function*, $K(.,.)$, as follows:

$$K(\mathbf{x}(t'), \mathbf{x}(t)) = \boldsymbol{\varphi}(\mathbf{x}(t'))^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}(t)) \quad t', t = 1, \ldots, N_{tr}. \tag{14}$$

In this paper, a Gaussian kernel is employed to replace the dot product in equation (14):

$$K(\mathbf{x}(t'), \mathbf{x}(t)) = \exp\left(-\frac{||\mathbf{x}(t') - \mathbf{x}(t)||^2}{\sigma^2}\right) \quad t', t = 1, \ldots, N_{tr}. \tag{15}$$

Note that equations (12) and (13) yield a set of linear equations, from which the machine learning parameters, b and $\boldsymbol{\alpha}$ are obtained for the given training set $\mathbf{D}_{tr}$:

$$\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^\mathsf{T} \\ \mathbf{1} & \Omega + \frac{1}{\gamma}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbf{P}_d \end{bmatrix} \tag{16}$$

where, $\mathbf{P}_d = [P_d(1), \ldots, P_d(N_{tr})]^\mathsf{T}$, $\mathbf{1} = [1, \ldots, 1]^\mathsf{T}$, $\mathbf{I}$ is the identity matrix, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N_{tr}}]^\mathsf{T}$, and the entries of the kernel matrix, Ω, are determined as follows:

$$\Omega_{t't} = K(\mathbf{x}(t'), \mathbf{x}(t)) \quad t', t = 1, \ldots, N_{tr}. \tag{17}$$

It should be noted that b and $\boldsymbol{\alpha}$ can have different values as the values of input parameters, n_{lag}, σ, and γ, change. To tune the regression model with respect to input parameters, k-fold cross-validation is performed. Moreover, mean absolute percentage error (MAPE) is adopted as the criteria for evaluating the performance of the regression model. After completing the cross-validation and training procedures, which optimize the input variables and learning parameters, the estimation accuracy of the regression model is evaluated on a test set, $\mathbf{D}_t$, of size N_t.

The critical step in calculating feeder CLPU demand ratio R_{CLPU} is to estimate the diversified feeder demand at the time of restoration, t_r. Similar to equation (8), the explanatory variable at time t_r is obtained as follows:

$$\mathbf{x}(t_r) = [P_d(t_r - 1), \ldots, P_d(t_r - n_{lag}^*), T(t_r)]^\mathsf{T} \tag{18}$$

where, n_{lag}^* is the optimal time lag. Using this explanatory variable, the estimated diversified feeder demand at time t_r is determined based on the trained LS-SVM model, as follows:

$$\hat{P}_d(t_r) = \sum_{t=1}^{N_{tr}} \alpha_t^* K(\mathbf{x}(t), \mathbf{x}(t_r)) + b^* \tag{19}$$

where, α_t^* and b^* are the optimal machine learning parameters.

Hence, R_{CLPU} is obtained by dividing the undiversified feeder restoration demand by the empirically averaged estimated diversified feeder demand at time t_r, as shown in equation (4). Note that the empirical averaging process is performed considering the estimation residual distribution obtained from a test set during normal system operation. An algorithmic overview of LS-SVM model for assessing R_{CLPU} is summarized in Algorithm 1.

IV. Customer Demand Increase Estimation

Although the aggregate residential demand at feeder level can be estimated with satisfactory accuracy, individual customer consumption can be quite stochastic [23]. Fig. 3 shows the daily demand curves of a single residential customer with and without outage, with 15-minute resolution during a season. The gray

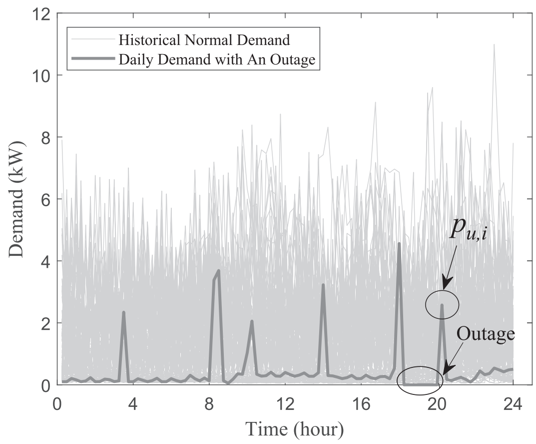


Fig. 3. Customer daily demand curves.

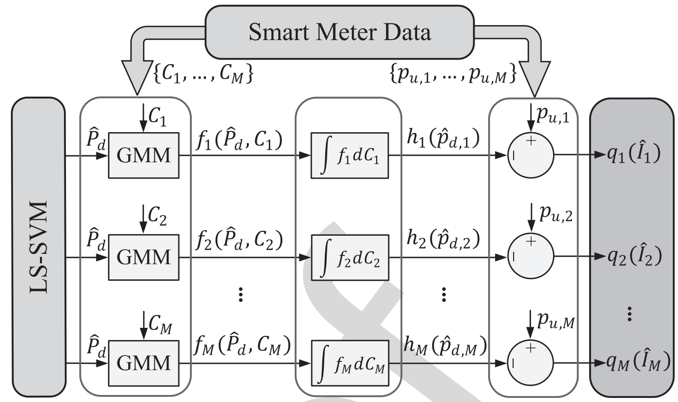


Fig. 4. Overall structure of the customer CLPU demand increase assessment.

Algorithm 1: LS-SVM.

1: Split demand and temperature data into two parts: training/validation set, and test set
2: **procedure** TRAINING/VALIDATION
3: Select initial n_{lag}, σ, γ
4: $\mathbf{D}_{tr} \leftarrow \{\mathbf{x}(t), P_d(t)\}_{t=1}^{N_{tr}}$
5: $\Omega \leftarrow K(\mathbf{x}(t'), \mathbf{x}(t)) \quad t', t = 1, \ldots, N_{tr}.$
6: $\mathbf{P}_d \leftarrow [P_d(1), \ldots, P_d(N_{tr})]^\mathsf{T}$
7: Solve equation (16) to obtain α_t's and b
8: $\hat{P}_d(t') \leftarrow \sum_{t=1}^{N_{tr}} \alpha_t K(\mathbf{x}(t), \mathbf{x}(t')) + b$
9: Compute validation MAPE
10: Change n_{lag}, σ, γ, do Step 3 to 9, optimize parameters, $n_{lag}^*, \sigma^*, \gamma^*, b^*, \boldsymbol{\alpha}^*$
11: **end procedure**
12: **procedure** TESTING $n_{lag}^*, \sigma^*, \gamma^*, b^*, \boldsymbol{\alpha}^*$
13: $\mathbf{D}_t \leftarrow \{\mathbf{x}(t), P_d(t)\}_{t=1}^{N_t}$
14: $\hat{P}_d(t) = \sum_{t'=1}^{N_{tr}} \alpha_{t'}^* K(\mathbf{x}(t'), \mathbf{x}(t)) + b^*$
15: Compute test MAPE
16: **end procedure**
17: **procedure** DEMAND ESTIMATION $(n_{lag}^*, \sigma^*, \gamma^*, b^*, \boldsymbol{\alpha}^*)$
18: $\mathbf{x}(t_r) \leftarrow [P_d(t_r - 1), \ldots, P_d(t_r - n_{lag}^*), T(t_r)]^\mathsf{T}$
19: $\hat{P}_d(t_r) \leftarrow \sum_{t=1}^{N_{tr}} \alpha_t^* K(\mathbf{x}(t), \mathbf{x}(t_r)) + b^*$
20: **end procedure**
21: $\hat{R}_{CLPU} \leftarrow P_u / E\{\hat{P}_d(t_r)\}$

curves are the historical daily demand data during normal operation and the red curve denotes the demand data in one day with an outage. The spikes represent the semi-periodic on/off cycling of customer appliances, which are captured by the smart meter. It can also be seen that customer demand at the interval [17.75 h, 20 h] equals zero, indicating an outage during this period. The plot shows that daily load curves do not present obvious cyclic behavior in contrast with the feeder demand. Also, customer peak demand at the time of restoration, $p_{u,i}$, is not necessarily larger than normal demand. Hence, considering the volatility of customer demand, in this section a probabilistic method is proposed to determine $\hat{I}_i$. The overall structure of customer-level CLPU demand assessment is shown in Fig. 4.

As discussed in Section II, the estimated demand for the i^{th} customer at the time of restoration is obtained as follows:

$$\hat{p}_{d,i}(t_r) = \hat{P}_d(t_r) C_i \quad i = 1, \ldots, M. \tag{20}$$

The difficulty in equation (20) is to compute the product of the two random variables $\hat{P}_d$ and C_i. It is shown in [24], that using the joint PDF of two dependent random variables, the marginal PDF of their product can be obtained using a nonlinear transformation. Hence, denoting the joint PDF of $\hat{P}_d$ and C_i by $f_i(\hat{P}_d, C_i)$, the marginal PDF of estimated customer demand $\hat{p}_{d,i}$ at time t_r is obtained as follows [24]:

$$h_i(\hat{p}_{d,i}) = \int_{0_+}^{1} f_i\left(\frac{\hat{p}_{d,i}}{C_i}, C_i\right) \frac{1}{C_i} dC_i \tag{21}$$

Therefore, the first step in calculating equation (21) is to obtain $f_i(\hat{P}_d, C_i)$ for each customer. To do this, a probabilistic technique is employed using GMMs. GMM is a parametric model, which approximates arbitrary PDFs as weighted summation of Gaussian density components. GMM has been previously applied in distribution systems studies for modeling the stochasticity of load and the uncertainty of distribution system state estimators [25], [26]. In this paper, we propose using GMM to model the joint PDF of customer contribution and the estimated diversified feeder demand. Thus, based on the estimated diversified feeder demand, $\hat{P}_d$, and contribution factor, C_i, for the i^{th} customer, the GMM approximation model, which is composed of S_i Gaussian components, can be expressed as follows:

$$l\left(\mathbf{z}|\boldsymbol{\lambda}\right) = \sum_{j=1}^{S_i} w_j g\left(\mathbf{z}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{22}$$

where, $\mathbf{z}$ is a two-dimensional continuous-valued vector defined as $\mathbf{z} = [\hat{P}_d, C_i]$, w_j's are the mixture weights corresponding to multi-variate Gaussian components $g(\mathbf{z}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, which satisfy $\sum_{j=1}^{S_i} w_j = 1$. Thus, each component is a bi-variate Gaussian function defined as:

$$g(\mathbf{z}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{z}-\boldsymbol{\mu}_j)\right\} \tag{23}$$

where, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the component mean vector and covariance matrix, respectively. Also, $\boldsymbol{\lambda} = \{w_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$, is the collection of parameters of the GMM model that have to be learned.

With respect to GMM approximation, the model training objective is to obtain the optimal parameter collection $\boldsymbol{\lambda}^*$ that best matches the distribution of the given training set. The training set is composed of C_i data history and $\hat{P}_d$ samples. The most well-established method for GMM training is the maximum likelihood (ML) estimation [27]. Given the training vectors $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ with N samples, the GMM likelihood can be written as:

$$l(Z|\boldsymbol{\lambda}) = \prod_{t=1}^{N} p(\mathbf{z}_t|\boldsymbol{\lambda}) \tag{24}$$

Generally, this non-linear function can be minimized iteratively with respect to $\boldsymbol{\lambda}$ using expectation-maximization (EM) algorithm [27]. Also, to tune the GMM models with respect to the component number, S_i, k-fold cross-validation is performed. To evaluate model performance with different learned $\boldsymbol{\lambda}$'s, the Bayesian information criterion (BIC) is employed.

Based on the optimal $\boldsymbol{\lambda}^*$ obtained from equation (24) and using equations (22) and (23), the joint PDF of estimated feeder demand $\hat{P}_d$ at time t_r and contribution factor C_i for the i^{th} customer can be specifically written as,

$$f_i(\hat{P}_d, C_i) = \sum_{j=1}^{S_i} \omega_j g(\hat{P}_d, C_i) \tag{25}$$

where,

$$g(\hat{P}_d, C_i) = \frac{1}{2\pi\sigma_{\hat{P}_d}^{(j)}\sigma_{C_i}^{(j)}\sqrt{1-\rho_j^2}} \exp\left(-\frac{1}{2(1-\rho_j^2)} \left[\frac{(\hat{P}_d-\mu_{\hat{P}_d}^{(j)})^2}{{\sigma_{\hat{P}_d}^{(j)}}^2} + \frac{(C_i-\mu_{C_i}^{(j)})^2}{{\sigma_{C_i}^{(j)}}^2} - \frac{2\rho_j(\hat{P}_d-\mu_{\hat{P}_d}^{(j)})(C_i-\mu_{C_i}^{(j)})}{\sigma_{\hat{P}_d}^{(j)}\sigma_{C_i}^{(j)}}\right]\right) \tag{26}$$

where, $\mu_{\hat{P}_d}^{(j)}, \mu_{C_i}^{(j)}, \sigma_{\hat{P}_d}^{(j)}, \sigma_{C_i}^{(j)}$, and ρ_j are the corresponding mean, variance, and correlation of $\hat{P}_d$ and C_i for the jth component, respectively. Hence, substituting equation (25) into (21), the marginal PDF of the estimated customer demand, $h_i(\hat{p}_{d,i})$, is obtained using numerical integration over the customer contribution factor variable.

Finally, using equation (7), the marginal PDF of demand increase for the i^{th} customer is constructed. Note that since $p_{u,i}$ is directly measured by the customer's smart meter at the time of restoration, it is treated as a constant value. Hence, the marginal PDF of $\hat{I}_i$, denoted as q_i is obtained for each customer, as follows:

$$q_i(\hat{I}_i) = h_i(p_{u,i} - \hat{I}_i) \tag{27}$$

TABLE I
OUTAGE CASE INFORMATION

Case	Outage Duration	Ambient Temperature
1	33 min	38.5 °C
2	91 min	37.0 °C
3	128 min	39.0 °C
4	136 min	35.0 °C
5	61 min	31.5 °C
6	83 min	34.0 °C
7	47 min	35.5 °C
8	101 min	30.0 °C
9	46 min	25.5 °C
10	77 min	29.5 °C
11	93 min	38.5 °C
12	41 min	31.0 °C
13	65 min	36.5 °C
14	48 min	29.0 °C
15	63 min	38.0 °C
16	118 min	26.5 °C
17	107 min	32.0 °C
18	85 min	27.0 °C
19	35 min	29.5 °C

V. Case Study

Nineteen outage cases are observed for evaluating feeder-level CLPU demand ratio and post-outage customer-level CLPU demand increase. The case information is shown in Table I.

A. *Feeder CLPU Demand Ratio Estimation*

1) CLPU Ratio Estimation and Regression Analysis: Feeder CLPU demand ratio is obtained by dividing the measured undiversified restoration demand by the estimated diversified demand at the time of restoration. As shown in Fig. 5, a demand overshoot occurs in the restoration phase, and the undiversified demand is significantly greater than the estimated diversified demand. Note that the undiversified CLPU demand (the spike labeled as P_u in Fig. 5) is observed using smart meter data, and is not estimated. Also, it is observed that once the restoration phase is completed, the actual feeder demand drops back to the estimated diversified levels. This corroborates the accuracy of the LS-SVM framework.

Table II shows the values of R_{CLPU} and the LS-SVM estimation MAPE. The performance of LS-SVM has been compared with two other regression models: 1) the autoregressive model with exogenous input variables (ARX), and 2) the polynomial NARX (P-NARX) model [28]. As is observed in Table II, LS-SVM shows better R_{CLPU} estimation accuracy compared with the other two models. From Table I and Table II, the impact

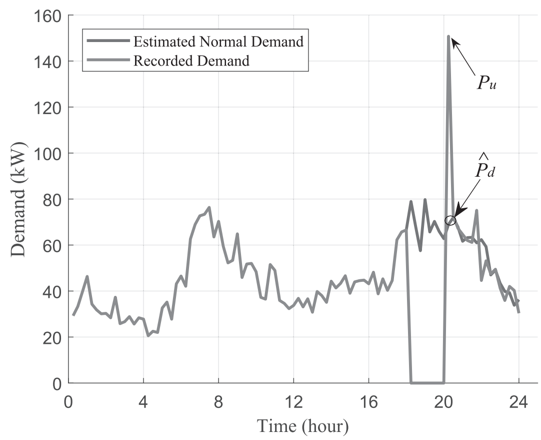


Fig. 5. Estimated diversified feeder demand curve and the recorded demand curve with an outage.

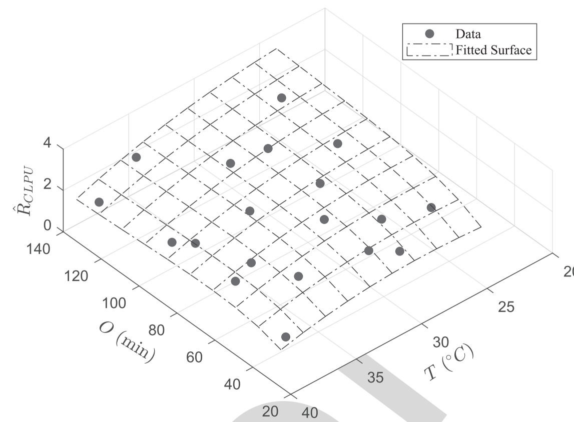


Fig. 6. Regression analysis of estimated CLPU ratios.

TABLE II
FEEDER CLPU DEMAND RATIO

Case	LS-SVM		P-NARX		ARX	
	Ratio	MAPE	Ratio	MAPE	Ratio	MAPE
1	1.41	5.45%	1.28	6.85%	1.37	6.38%
2	1.63	10.89%	1.56	13.59%	1.55	13.70%
3	1.88	11.56%	1.83	13.70%	1.81	13.55%
4	2.16	10.77%	2.25	12.30%	2.19	11.88%
5	2.87	10.48%	2.97	13.53%	2.88	10.45%
6	2.69	7.09%	2.72	5.62%	2.66	8.25%
7	2.40	8.51%	2.38	8.60%	2.34	6.80%
8	3.17	10.10%	3.31	11.42%	3.20	11.07%
9	2.71	10.85%	2.82	14.33%	2.79	13.47%
10	2.88	7.14%	2.95	9.53%	2.99	9.28%
11	2.06	9.52%	2.04	10.17%	2.00	13.85%
12	2.50	13.19%	2.54	14.43%	2.51	13.45%
13	2.22	5.21%	2.25	6.50%	2.21	5.08%
14	2.88	7.50%	2.94	9.53%	2.93	9.28%
15	1.97	9.18%	1.95	9.22%	1.98	9.17%
16	3.31	8.35%	3.35	11.07%	3.39	11.74%
17	2.74	5.07%	2.79	3.40%	2.74	5.16%
18	3.43	5.87%	3.50	8.83%	3.47	7.78%
19	2.35	9.74%	2.43	13.06%	2.42	10.93%

of outage duration and ambient temperature on the ratio can be observed.

Fig. 6 shows the regression analysis result of the estimated CLPU ratio in terms of outage duration (O) and ambient temperature (T). As can be seen, a surface is fitted to the data with acceptable accuracy using polynomial regression based on the estimated CLPU ratios. This CLPU ratio regression model provides an alternative way for estimating the CPLU ratio and demand in future system restoration cases. Also, as more outage cases are collected, the accuracy of the CLPU ratio regression model can be improved.

2) *Model Robustness Evaluation:* The robustness of learning parameters (σ and γ) has been tested in accordance with [29]. To do this, the following steps have been performed: first, 1% of demand samples in the training set are randomly selected. Second, the selected samples are contaminated by multiplying them with different contamination coefficients to generate outliers. The contamination coefficient is varied from 1 to 2 with step of 0.1. The contaminated demand outliers can be written as follows:

$$P_{ou} = P_{or} K_m \tag{28}$$

where, P_{or} is the original uncontaminated demand sample and K_m denotes the contamination coefficient. Third, for each contamination coefficient, the LS-SVM model is retrained to obtain the new learning parameters. Also, the MAPE under these new parameters is obtained over the test set. Finally, the ratios of the retrained learning parameters (with outliers) to the original learning parameters (without outliers) are calculated to quantify the changes in the model due to noise injection. These ratios are also obtained by dividing the estimation MAPEs of the model with and without outliers. The ratios are written as follows:

$$K_\sigma = \frac{\sigma_{ou}}{\sigma_{or}} \times 100\ \% \tag{29}$$

$$K_\gamma = \frac{\gamma_{ou}}{\gamma_{or}} \times 100\ \% \tag{30}$$

$$K_{MAPE} = \frac{MAPE_{ou}}{MAPE_{or}} \times 100\ \% \tag{31}$$

where, σ_{ou}, γ_{ou} denote the retrained learning parameters after contamination, $MAPE_{ou}$ denotes the estimation MAPE corresponding to σ_{ou} and γ_{ou}; similarly, σ_{or}, γ_{or} denote the trained learning parameters obtained from the original uncontaminated training set, and $MAPE_{or}$ corresponds to σ_{or} and γ_{or}. Fig. 7 shows the changes in learning parameters and test MAPE against the contamination coefficient. It can be seen that the estimation MAPE has been kept unchanged, since the LS-SVM has been able to automatically adjust itself to higher levels of noise to maintain satisfactory performance. In addition, parameter σ has been nearly kept constant as the contamination coefficient increases. However, parameter γ decreases as the contamination

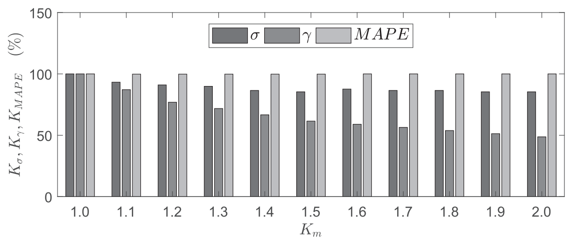
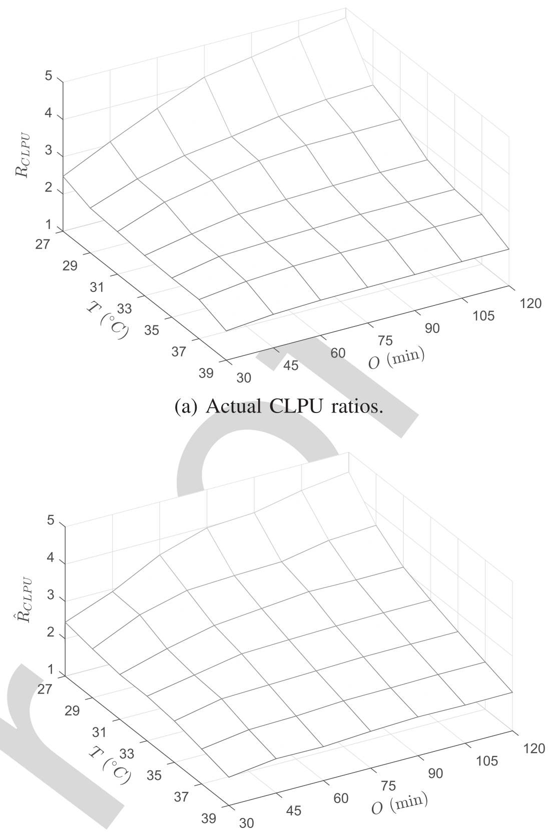


Fig. 7. Robustness evaluation of learning parameters.



(a) Actual CLPU ratios.

(b) Estimated CLPU ratios.

Fig. 8. Actual and estimated CLPU ratios.

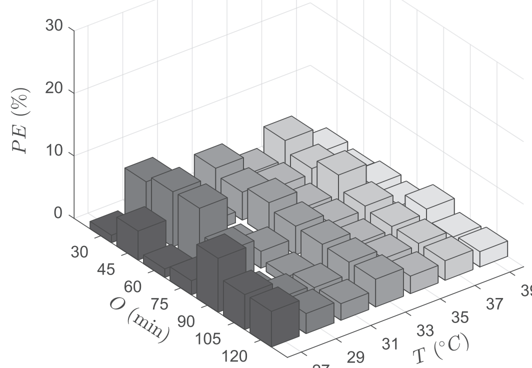


Fig. 9. CLPU ratio estimation percentage errors.

coefficient increases, which can be justified using equation (10). As can be seen, in equation (10), γ is basically the regularization trade-off factor which determines the weight of the training set's inherent noise level during the training process. Hence, when we artificially increase this noise level in the training set through contamination, the LS-SVM training algorithm automatically decreases the weight assigned to the inherent noise parameter to keep the risk function at its minimum value.

3) CLPU Ratio Validation: Since the real diversified demand at the time of restoration is unknown, due to the undiversified nature of load, therefore, we cannot validate the estimated ratios using real outage cases alone. Considering this, we have conducted additional Monte Carlo simulations to validate the CLPU ratios. Specifically, 49 new outage cases with different outage durations and different ambient temperatures are created, and then, our proposed approach is applied to the data generated from these outage cases [30]. The basic steps of validation are as follows: first, the demand consumed by TCLs are generated using Monte Carlo simulations for a heterogeneous population of customers, obtained from appliance-level state-space modeling, both in normal operation and outage conditions; second, additional appliances' consumed demands are added to the TCLs' consumed demands to obtain customers' net demands; third, feeder-level demand is obtained by aggregating customer-level loads; then, the proposed CLPU assessment framework is applied to the data obtained from 49 outage cases to obtain corresponding CLPU ratios; finally, ratio validation is conducted by comparing the outcomes of the proposed data-driven model and the simulation results. Fig. 8(a) and Fig. 8(b) show the actual and estimated CLPU ratios, respectively. Note that the actual CLPU ratios are obtained from Monte Carlo simulations, and the estimated CLPU ratios are obtained by applying our proposed framework to the demand data generated from these Monte Carlo simulations. In Fig. 8(a) and Fig. 8(b), T denotes the ambient temperature and O denotes outage duration. As can be seen, the estimated CLPU ratios can accurately match the actual CLPU ratios. The validation of CLPU ratio can also be demonstrated in Fig. 9, in which the CLPU ratio estimation percentage errors (PE) are smaller than 9% for all cases, and 90% of the percentage error values are less than 6%, which validates the performance of the framework. This can also demonstrate the advantage of our proposed data-driven approach over the model-driven Monte Carlo simulator, showing that the CLPU ratio can be accurately estimated only based on the available demand data and without the knowledge of thermal parameters of individual customer houses.

In practice, it is probable that a proportion of customers are unmonitored. Hence, it is of interest to analyze the performance of the proposed framework in scenarios where different proportions of customers do not have smart meters. This has also been demonstrated using Monte Carol simulations, where the CLPU ratio estimation percentage errors are shown as a function of percentage of monitored customers in Fig. 10. It can be seen that as the number of monitored customers increases the accuracy of the proposed framework improves. Also, the framework still has

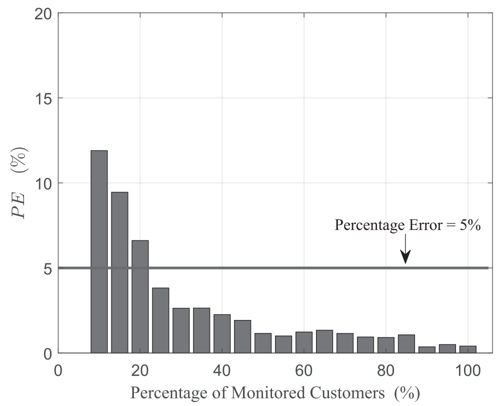


Fig. 10. Relationship between the CLPU ratio estimation percentage error with the percentage of monitored customers.

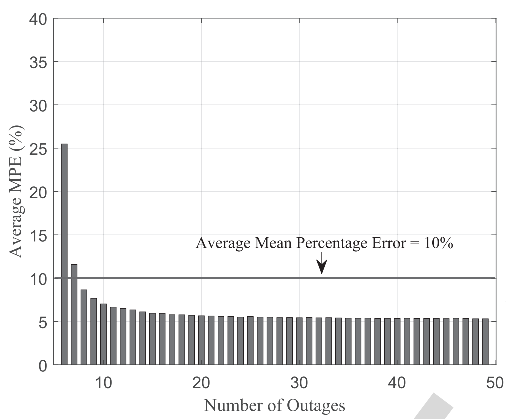


Fig. 11. Relationship between average MPE with the number of outages.

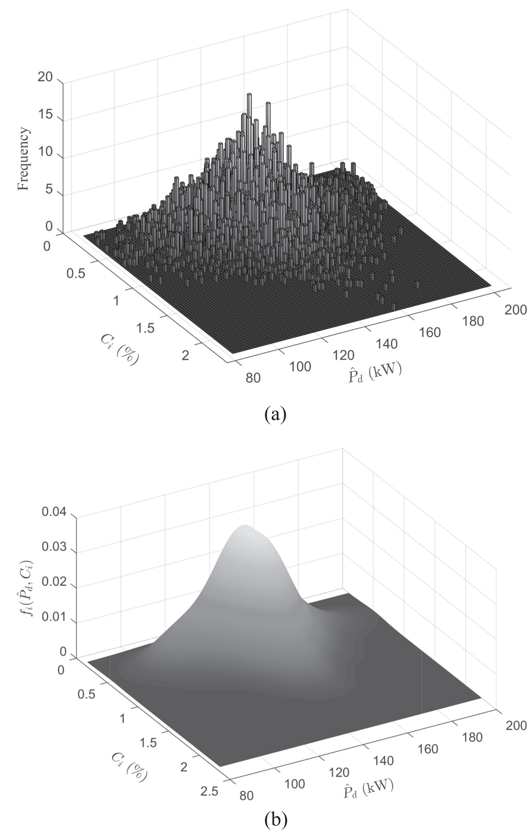


Fig. 12. Joint PDF estimation of diversified feeder demand and contribution factor for one customer. (a) Empirical histogram. (b) GMM-based estimation.

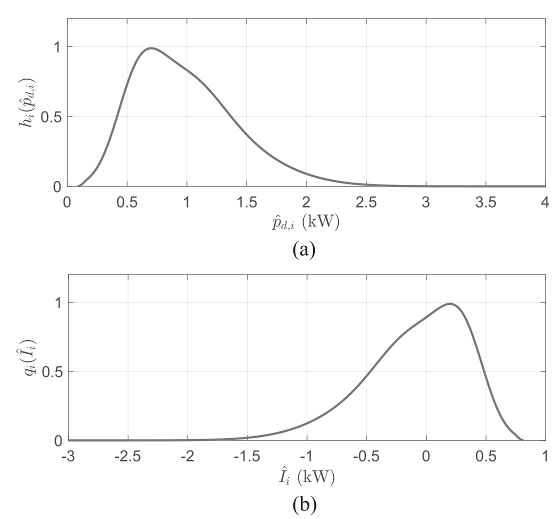


Fig. 13. Distribution of estimated demand and CLPU demand increase of one customer. (a) Distribution of $\hat{p}_{d,i}$. (b) Distribution of $\hat{I}_i$.

acceptable accuracy even when a high percentage of customers are unmonitored.

It is also of great importance to conduct robustness analysis with respect to the number of available outage cases due to the outage data scarcity. To do this, a training process has been performed using random drop-out for cross validation. The performance of the framework has been evaluated in terms of the CLPU ratio prediction mean percentage error (MPE), and is plotted against the number of historical outages, as shown in Fig. 11. As can be seen, to reach an average MPE of smaller than 10%, a minimum number of eight outages is required in this case. Hence, as more outage data become available, the accuracy of the regression model is improved. This robustness analysis has also been conducted on our utility data, and a similar decreasing trend of average MPE against the number of outages is observed.

B. Customer CLPU Demand Increase Estimation

Fig. 12(a) and Fig. 12(b) show the empirical histogram and the GMM-based estimation of $f_i(\hat{P}_d, C_i)$ for one customer, respectively. As can be seen by comparing these figures, GMM is able to accurately model the behavior of the customer using smooth parametric Gaussian density functions.

Fig. 13 shows the probability distribution of estimated demand and CLPU demand increase of one customer at time t_r. Note that the probable CLPU demand increase of the customer

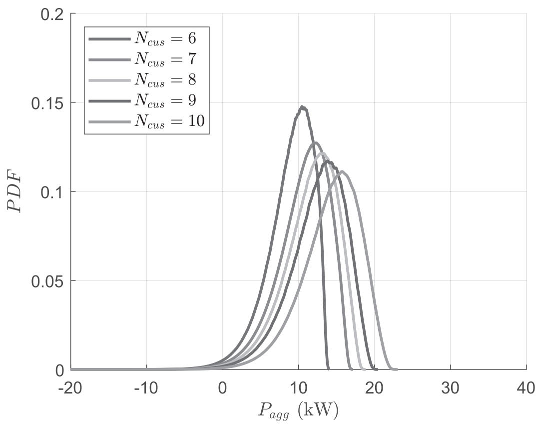


Fig. 14. Distributions of aggregate demand increase.

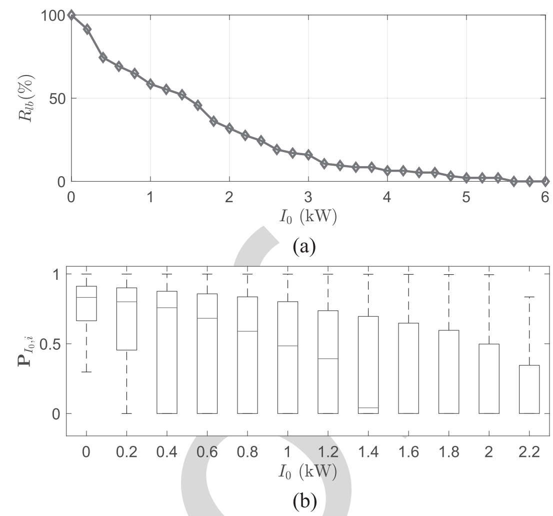


Fig. 15. Evaluation of loss of load diversity. (a) The relationship between R_{lb} and I_0. (b) Distribution of $\mathbb{P}_{I_0,i}$.

can be negative. This partly reflects the stochasticity of customer demand. Regarding system restoration issue, optimal approaches have been proposed for restoring different groups of customers after extreme events in the literature [6]–[8]. Our proposed method can provide the marginal PDF of demand increase for a group of customers by convolving the marginal PDFs of demand increase of individual customers [31], which is useful for the utilities to perform restoration risk evaluation. For instance, Fig. 14 shows the PDFs of aggregate demand increase (P_{agg}) for N_{cus} customers connected to the same transformer. Hence, the impact of CLPU demand increase on the transformer can be accurately quantified. As can be seen, as the number of customers increases the expected aggregate demand increase also shifts towards larger values.

To evaluate the loss of load diversity for a population of customers, the following index is defined for each customer:

$$\mathbb{P}_{I_0,i} = \Pr(\hat{I}_i \geq I_0) \tag{32}$$

where, $\mathbb{P}_{I_0,i}$ denotes the probability of estimated demand increase being larger than a threshold, I_0, for the i^{th} customer, with $\Pr(a)$ defining probability of event a. Using this index, the factor $R_{lb}(I_0)$ indicates the percentage of customers with $\mathbb{P}_{I_0,i} > 0$, as shown in equation (33):

$$R_{lb}(I_0) = \frac{\sum_{i=1}^{M} H(\mathbb{P}_{I_0,i})}{M} \times 100\% \tag{33}$$

where, $H(x)$ is the Heaviside step function defined as follows:

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{34}$$

Fig. 15(a) shows the relationship between R_{lb} and I_0. It can be seen that: 1) for $I_0 = 0$ we have $R_{lb} = 100\%$, which implies that all customers have non-negative CLPU demand increase with non-zero probability, and 2) R_{lb} decreases as I_0 increases, which indicates that the number of customers with $\hat{I}_i > I_0$ decreases as I_0 increases. This is determined by the maximum capability of customers' contribution to feeder CLPU demand. Fig. 15(b) shows $\mathbb{P}_{I_0,i}$ distribution boxplot as a function of threshold level I_0. This figure describes the loss of load diversity during service restoration. For example, the first box tells us that almost all of the customers have $\mathbb{P}_{I_0} = 0, i > 0.5$ for this outage case. This means that nearly all customer' loads simultaneously start drawing more energy than normal from the feeder in the restoration phase. It can also be seen that the first quartile, the median, and the third quartile values of $\mathbb{P}_{I_0,i}$ present a descending trend as the threshold I_0 increases. This is consistent with the decreasing trend of R_{lb}, observed in Fig. 15(a). This implies that only a few customers have abnormaly high demand increase during service restoration.

It is also of interest to discover the relationship between the uncertainty of customer demand increase and the uncertainty of customer behavior during normal system operation. To evaluate the uncertainty of customer demand increase at the time of restoration, the *entropy* of $\hat{I}_i$ is obtained using $q_i(\hat{I}_i)$, as follows [32]:

$$E(\hat{I}_i) = -\int_{\hat{I}_i} q_i(\hat{I}_i) \log_2(q_i(\hat{I}_i)) d\hat{I}_i \tag{35}$$

On the other hand, to evaluate the uncertainty of customer behavior during normal system operation, customer demand is sampled on different days at the same time corresponding to the restoration instant. Based on these data samples the entropy of customer behavior is defined similar to (35) and denoted as $E(p_{d,i})$. Fig. 16 shows the relationship between $E(\hat{I}_i)$ and $E(p_{d,i})$ for all customers. It can be seen that a positive linear relationship exists between the uncertainty of customer CLPU demand increase and the uncertainty of normal customer demand at the time corresponding to the restoration instant. The correlation between these two entropy variables is around 0.72, which implies that customers with uncertain normal demand also show more uncertainty at the time of restoration.

To assess customer demand increase due to CLPU, the relationship between energy consumption within a 4-hour time

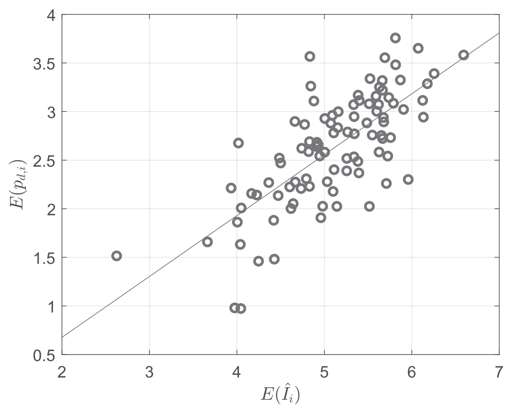


Fig. 16. Relationship between entropy of customer CLPU demand increase and entropy of normal demand at the time corresponding to restoration instant.

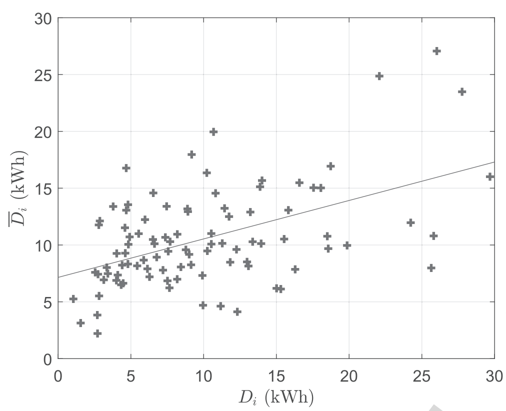


Fig. 17. Relationship between customer energy consumption after the time of restoration and average normal energy consumption at corresponding time period.

interval after time of restoration, D_i, and average energy consumption during the same time period in normal operation, $\overline{D}_i$, is analyzed. Fig. 17 shows the relationship between D_i and $\overline{D}_i$ for all customers. The slope of the fitted line is smaller than 1, which indicates that customer energy consumption after the time of restoration is greater than average energy consumption in normal operation during the corresponding time period. Also, a positive correlation between D_i and $\overline{D}_i$ is observed, which implies that higher energy consumption during normal system operation corresponds to higher restoration energy consumption.

VI. Conclusion

This paper has presented a data-driven framework for using smart meter data to determine feeder-level CLPU demand ratio and to assess customer-level demand increase due to CLPU, based on historical outage cases. Machine learning and probabilistic methodologies are used for CLPU demand assessment. Outage cases are employed for model training and verification. The results of case studies show that the proposed framework can accurately determine feeder-level CLPU demand ratio and assess customer-level demand increase due to loss of load diversity during service restoration. It is shown that only a few customers have extreme CLPU demand increase, and customers with higher energy consumption during normal operation typically have higher demand during the restoration phase. The performance of the proposed data-driven framework is validated using extensive Monte Carlo simulations. It has been demonstrated that our method is able to accurately assess CLPU demand at both feeder- and customer-levels without having any explicit knowledge of individual houses' thermal information.

References

[1] D. Athow and J. Law, "Development and applications of a random variable model for cold load pickup," *IEEE Trans. Power Del.*, vol. 9, no. 3, pp. 1647–1653, Jul. 1994.

[2] W. W. Lang, M. D. Anderson, and D. R. Fannin, "An analytical method for quantifying the electrical space heating component of a cold load pick up," *IEEE Trans. Power App. Syst.*, vol. PAS-101, no. 4, pp. 924–932, Apr. 1982.

[3] K. P. Schneider, E. Sortomme, S. S. Venkata, M. T. Miller, and L. Ponder, "Evaluating the magnitude and duration of cold load pick-up on residential distribution using multi-state load models," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3765–3774, Sep. 2016.

[4] V. Kumar, R. Kumar H. C., I. Gupta, and H. O. Gupta, "DG integrated approach for service restoration under cold load pickup," *IEEE Trans. Power Del.*, vol. 25, no. 1, pp. 398–406, Jan. 2010.

[5] A. Al-Nujaimi, M. A. Abido, and M. Al-Muhaini, "Distribution power system reliability assessment considering cold load pickup events," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4197–4206, Jul. 2018.

[6] B. Chen, C. Chen, J. Wang, and K. L. Butler-Purry, "Multi-time step service restoration for advanced distribution systems and microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6793–6805, Nov. 2018.

[7] A. Arif *et al.*, "Optimizing service restoration in distribution systems with uncertain repair time and demand," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6828–6838, Nov. 2018.

[8] J. C. Lpez, J. F. Franco, M. J. Rider, and R. Romero, "Optimal restoration/maintenance switching sequence of unbalanced three-phase distribution systems," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6058–6068, Nov. 2018.

[9] J. Aubin, R. Bergeron, and R. Morin, "Distribution transformer overloading capability under cold-load pickup conditions," *IEEE Trans. Power Del.*, vol. 5, no. 4, pp. 1883–1891, Oct. 1990.

[10] F. Edstrom, J. Rosenlind, K. Alvehag, P. Hilber, and L. Soder, "Influence of ambient temperature on transformer overloading during cold load pickup," *IEEE Trans. Power Del.*, vol. 28, no. 1, pp. 153–161, Jan. 2013.

[11] V. Gupta and A. Pahwa, "A voltage drop-based approach to include cold load pickup in design of distribution systems," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 957–963, May 2004.

[12] J. J. Wakileh and A. Pahwa, "Optimization of distribution system design to accommodate cold load pickup," *IEEE Trans. Power Del.*, vol. 12, no. 1, pp. 339–345, Jan. 1997.

[13] R. E. Mortensen and K. P. Haggerty, "A stochastic computer model for heating and cooling loads," *IEEE Trans. Power Syst.*, vol. 3, no. 3, pp. 1213–1219, Aug. 1988.

[14] C. Chong and R. P. Malhami, "Statistical synthesis of physically based load models with applications to cold load pickup," *IEEE Power Eng. Rev.*, vol. PER-4, no. 7, pp. 33–33, Jul. 1984.

[15] K. P. Schneider, J. C. Fuller, and D. P. Chassin, "Multi-state load models for distribution system analysis," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2425–2433, Nov. 2011.

[16] E. Agneholm and J. Daalder, "Cold load pick-up of residential load," *IEE Proc.—Gener., Transmiss. Distrib.*, vol. 147, no. 1, pp. 44–50, Jan. 2000.

[17] 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*, vol. 10, no. 2, pp. 2312–2322, Mar. 2019.

[18] M. Espinoza, J. A. K. Suykens, R. Belmans, and B. D. Moor, "Electric load forecasting: Using kernel based modeling for nonlinear system identification," *IEEE Control Syst. Mag.*, vol. 27, no. 5, pp. 43–57, Oct. 2007. Q2

[19] National Oceanic and Atmospheric Administration. [Online]. Available: https://www.noaa.gov/

[20] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Sci., 2002.

[21] Z. S. Hosseini, M. Mahoor, and A. Khodaei, "Ami-enabled distribution network line outage identification via multi-label SVM," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5470–5472, Sep. 2018.

[22] H. Jiang, Y. Zhang, E. Muljadi, J. J. Zhang, and D. W. Gao, "A short-term and high-resolution distribution system load forecasting approach using support vector regression with hybrid parameters optimization," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3341–3350, Jul. 2018.

[23] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[24] A. G. Glen, L. M. Leemis, and J. H. Drew, "Computing the distribution of the product of two continuous random variables," *Comput. Statist. Data Anal.*, vol. 44, no. 3, pp. 1–14, Jul. 2002.

[25] R. Singh, B. C. Pal, and R. A. Jabr, "Statistical representation of distribution system loads using Gaussian mixture model," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 29–37, Feb. 2010.

[26] G. Valverde, A. T. Saric, and V. Terzija, "Stochastic monitoring of distribution networks including correlated input variables," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 246–255, Feb. 2013.

[27] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, 2nd ed. New York, NY, USA: Springer, 2015, pp. 827–832.

[28] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, Spatio-Temporal Domains*. Hoboken, NJ, USA: Wiley, 2013.

[29] J. Suykens, J. D. Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomputing*, vol. 48, no. 1, pp. 85–105, Jun. 2002.

[30] S. Bashash and H. K. Fathy, "Modeling and control of aggregate air conditioning loads for robust renewable power management," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 4, pp. 1318–1327, Jan. 2013.

[31] V. V. Petrov, *Sums of Independent Random Variables*. New York, NY, USA: Springer-Verlag, 1975.

[32] J. Beirlant, E. Dudewicz, L. Gyorf, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Statist. Sci.*, vol. 6, pp. 17–39, Jul. 2001.

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 was a Commissioning Engineer for NARI Technology Co., Ltd., Nanjing, China. From 2013 to 2017, he was an Electrical Engineer with the State Grid Corporation of China, Jiangsu, Nanjing, China. His research interests include load modeling, load forecasting, data analytics in distribution system, and power system relaying.

Kaveh Dehghanpour (S'14–M'17) received the B.Sc. and M.S. degrees from the University of Tehran, Tehran, Iran, in electrical and computer engineering, in 2011 and 2013, respectively. He received 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.

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 with 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 data-driven 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 is the Secretary of IEEE Power and Energy Society Award Subcommittee. He is an Editor of the IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, and IEEE PES LETTERS, and an Associate Editor of *IET Smart Grid*.

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