Enriching Load Data Using Micro-PMUs and Smart Meters

Fankun Bu, Graduate Student Member, IEEE, Kaveh Dehghanpour, Member, IEEE, and Zhaoyu Wang, Senior Member, IEEE

Abstract—In modern distribution systems, load uncertainty can be fully captured by micro-PMUs, which can record high-resolution data; however, in practice, micro-PMUs are installed at limited locations in distribution networks due to budgetary constraints. In contrast, smart meters are widely deployed but can only measure relatively low-resolution energy consumption, which cannot sufficiently reflect the actual instantaneous load volatility within each sampling interval. In this paper, we have proposed a novel approach for enriching load data for service transformers that only have low-resolution smart meters. The key to our approach is to statistically recover the high-resolution load data, which is masked by the low-resolution data, using trained probabilistic models of service transformers that have both high- and low-resolution data sources, i.e., micro-PMUs and smart meters. The overall framework consists of two steps: first, for the transformers with micro-PMUs, a Gaussian Process is leveraged to capture the relationship between the maximum/minimum load and average load within each low-resolution sampling interval of smart meters; a Markov chain model is employed to characterize the transition probability of known high-resolution load. Next, the trained models are used as teachers for the transformers with only smart meters to decompose known low-resolution load data into targeted high-resolution load data. The enriched data can recover instantaneous load uncertainty and significantly enhance distribution system observability and situational awareness. We have verified the proposed approach using real high- and low-resolution load data.

Index Terms—Distribution system, load uncertainty, micro-PMU, smart meter, data enrichment.

I. INTRODUCTION

A S THE advanced metering infrastructure (AMI) has been widely deployed in distribution systems in recent years, utilities have gained access to large amounts of smart meter (SM) data [1]. To take advantage of this data, which is both spatially and temporally fine-grained, researchers and industry practitioners have performed time-series power flow studies for optimizing network operation, expansion [2], [3], and integrating renewable energy resources [4]. In many cases, customer-level demands are aggregated to obtain service transformer-level loads for performing power flow studies [2], [5]. However, the problem is that in most cases, SMs have a low sampling rate, e.g., one to four samples per hour. Thus, the average demand measured at such low resolutions cannot faithfully represent the uncertainties of the instantaneous load. As illustrated in Fig. 1 for an exemplary transformer, the maximum 1-second load data has reached values 40% times larger than the corresponding hourly SM reading within the same sampling interval. Also, compared to the hourly measurements, the instantaneous load shows a high level of variability, which has not been captured by the SMs. Therefore, recovering the masked high-resolution load data is critical in enhancing distribution system situational awareness and granularity of modeling.

To further demonstrate the usefulness of high-resolution load data, we primarily focus on three specific applications. First, accurate power flow analysis requires high-resolution load data. Power flow analysis is critically important for utilities. It can provide voltage profiles, which can help utilities plan new circuits, add customers, and track and fix voltage problems. Since load is an essential component in distribution systems, high-resolution load profiles play a critical role in obtaining power flow solutions with high fidelity. In contrast, the 15-min, 30-min, or 1-hour load data might cause unacceptable errors [6], [7]. This is why most utilities take conservative approaches in distribution system operation and planning. Instead, taking full advantage of high-resolution load data can free utilities from conservative measures. Second, accurate voltage regulation analysis requires high-resolution

Fig. 1. A one-day real service transformer load curve with 1-second load data and the corresponding hourly average load curve.
load data. In many cases, utilities perform time-series power flow analysis to examine the actions of voltage regulation devices. Typical voltage regulation devices include voltage regulators and capacitors. The controller of these two types of devices usually has a time delay before executing a regulating order. By doing this, the voltage regulation devices can avoid unnecessary frequent reactions to fast and temporary voltage transients. The time delay is typically around 30 seconds. To accurately capture the response of voltage regulation devices, the time resolution of load data for performing time-series power flow analysis should match the delay time of regulation devices’ controller [6], [8]. Third, high-resolution load data can facilitate photovoltaic (PV) integration. In most cases, utilities conservatively maintain customer voltages very close to the upper bound of the ANSI voltage range due to conservative considerations. Under this condition, even though the load increase may cause a voltage drop, the voltage will still be within the ANSI voltage range and satisfy voltage quality requirements. However, under such conservative operation logics, new PV integration can cause over-voltages. To assess the impact of PV generation, one promising way is to utilize high-resolution (1-second or 1-min) PV generation data to perform power flow analysis, because low-resolution data might fail to capture PV output variations. Since the load variations might not be negligible in some scenarios, it is necessary to combine high-resolution load data and PV output data to perform time-series power flow analysis [9], [10].

There is only a limited number of previous works focusing on load data enrichment. In [8], a top-down method is presented to generate service transformer-level high-resolution load profiles. First, low-resolution substation load profiles are allocated to service transformers via scaling. Then, the allocated profiles are decomposed into high-resolution load data by aggregating typical load patterns stored in variability and diversity libraries. In [11], synthetic load datasets are created for four typical seasonal months using captured variability from high-resolution service transformer load data. To develop rich load data, researchers have added random noise to load data for modeling load uncertainty, as presented in [12]. In [13], a discrete wavelet transform (DWT)-based approach is proposed to parameterize intra-second variability of high-resolution transformer load data. To sum up, the primary limitations of previous load data enrichment methods are: the scaled substation load profiles allocated to service transformers differ from the actual load profiles since each transformer has a distinct load pattern [14], inaccuracy of adding random noise, and lack of specific methodology for applying the extracted load variability [7].

Considering the shortcomings of previous works, in this paper, we have developed a novel bottom-up approach for enriching hourly load data for service transformers that only have SMs, by leveraging the high-resolution load data of service transformers with micro-PMUs and SMs. This concept is illustrated in Fig. 2, where the service transformer in the middle with rich load data is utilized to perform load data enrichment for the other two service transformers with only SMs. Before proceeding to specific steps, we have observed that each low-resolution load observation corresponds to a segment of high-resolution load profile, as shown in Fig. 1. Therefore, enriching one known low-resolution load observation comes down to determining the maximum and minimum loads in the corresponding high-resolution load profile segment and inferring how the instantaneous load varies within those bounds. To do this, the proposed approach exploits learned probabilistic models that are trained using the high-resolution load data of service transformers with micro-PMUs. Thus, the first stage is to train probabilistic models using known high-resolution load data of micro-PMUs. Specifically, a Gaussian Process is used to capture the relationship between the maximum/minimum bound and the average load. A Markov process is leveraged to model the probabilistic transition of instantaneous load within the bounds. These trained models for transformers with micro-PMUs form a teacher repository. The second stage is to extend the trained probabilistic models to the service transformers that only have SMs, i.e., the students, for enriching low-resolution load data. Specifically, the trained Gaussian Process models are employed to estimate the unknown maximum/minimum bound using the known low-resolution observation as the input, and the trained Markov models are used to probabilistically determine the variability of instantaneous load within the estimated maximum and minimum bounds. In addition, the load enrichment process in the second stage is performed using a weighted averaging operation, where the weights are determined by evaluating the similarity between low-resolution load data of the student and teacher transformers. Our approach is not restricted to the condition that the teacher and student transformers should have the same rating, loss, or served customer number. The overall framework of our proposed approach is illustrated in Fig. 3.

The primary contribution of our paper is that we have proposed a novel bottom-up inter-service-transformer load data enrichment approach using micro-PMUs and SMs. Our method takes full advantage of the fine-grained spatial and temporal granularity of SM and micro-PMU data. The rest of the paper is organized as follows: Section II presents the process of training teacher models using data from transformers with micro-PMUs. Section III describes the procedure of enriching load data for transformers with only SMs using
the uncertainty of functions evaluated at $P_a(t)$. In GPR, the function $f(P_a(t))$ is distributed as a Gaussian process:

$$f(P_a(t)) \sim \mathcal{GP}(\mu(P_a(t)), K(P_a(t), P_a(t'))),$$

Where, $\mu(P_a(t))$ reflects the expected value of the maximum load inference function, and the covariance function $K(P_a(t), P_a(t'))$ represents the dependence between the maximum loads during different hour intervals. In our problem, the covariance function $K(\cdot, \cdot)$ is specified by the Squared Exponential Kernel function expressed as:

$$K(P_a(t), P_a(t')) = \sigma^2 \exp \left( -\frac{|t_a(t) - t_a(t')|^2}{2\lambda^2} \right),$$

Where, $|\cdot|^2$ represents $l_2$-norm, $\sigma$ and $\lambda$ are hyper-parameters, which are determined using cross-validation. Intuitively, (3) measures the distance between $P_a(t)$ and $P_a(t')$, which can also reflect the similarity between $P(t)$ and $P(t')$, as shown in Fig. 4. For each service transformer with a micro-PMU, the average load and corresponding maximum load during each hour interval are known and provide a training dataset. Thus, applying (2) to the entire training dataset consisting of $N$ hourly average and maximum load pairs, $\{(P_a(1), \overline{P}(1)), \ldots, (P_a(N), \overline{P}(N))\}$, an N-dimensional joint Gaussian distribution can be constructed as:

$$\begin{bmatrix}
\tilde{f}(P_a(1)) \\
\vdots \\
\tilde{f}(P_a(N))
\end{bmatrix} \sim \mathcal{N}(\mu, \Sigma),$$

Where,

$$\mu = \begin{bmatrix}
\mu(P_a(1)) \\
\vdots \\
\mu(P_a(N))
\end{bmatrix},\quad \Sigma = \begin{bmatrix}
K(P_a(1), P_a(1)) & \cdots & K(P_a(1), P_a(N)) \\
\vdots & \ddots & \vdots \\
K(P_a(N), P_a(1)) & \cdots & K(P_a(N), P_a(N))
\end{bmatrix}.$$

The joint Gaussian distribution formulated in (4) represents a trained nonparametric maximum load inference model. Also, the same procedure can be applied to the hourly average and minimum load pairs, $\{(P_a(1), \underline{P}(1)), \ldots, (P_a(N), \underline{P}(N))\}$, to obtain a trained nonparametric minimum load inference model.

In summary, for each service transformer with a micro-PMU, we can obtain two trained GPR models for inferring the maximum and minimum loads based on the corresponding hourly average load measured at the low-resolution sampling intervals.

### B. Training Load Variability Inference Model

Given an hourly average load observation, simply determining load boundaries is not sufficient for load data enrichment. We also have to learn how the load varies within these bounds. It is observed from real high-resolution load data that when an appliance is turned on, the load will jump to a certain level, as shown in Fig. 5. This process can be modeled as the Markov chain, which represents a system transitioning from one state...
to another over time. Also, it is observed from Fig. 5 that once
the load has transitioned to a certain level, it will stay almost
invariant for a certain period of time. Therefore, the load state
duration demonstrates statistical properties, and there exists a
temporal correlation in state transition. Considering this, we
have employed a second-order Markov model to capture the
stochastic variability of the instantaneous load. Markov chains
of second order are processes in which the next state depends
on two preceding ones.

Since load is continuous, the first step to parameterize a
Markov chain process is to discretize high-resolution load
measurements. Specifically, for the th high-resolution load
observation during the tth hour interval, the corresponding
observed state is determined as:

\[ S_t(i) = n_s, \quad n_s \in \{1, \ldots, N_s\}, \quad t = 1, \ldots, N, \]
\[ \text{if} \quad (n_s - 1) \frac{P(t) - \bar{P}(t)}{N_s} \leq P_t(i) - \bar{P}(t) < n_s \frac{P(t) - \bar{P}(t)}{N_s}, \]

where, \( N_s \) represents the total number of the unique discrete
states and \( P_t(i) \) is the ith instantaneous load measurement
during the tth hour.

Also, it is observed from real high-resolution load data
that different load levels display different stochastic processes.
Typically, an air-conditioner cyclically starts and stops in
the order of minutes. In contrast, the baseload, which is
often caused by lighting and electronic devices, shows signif-
ically longer cycles. In addition, the air-conditioning devices
and baseload appliances show different average load levels
over low-resolution sampling intervals due to different capaci-
ties. Therefore, to capture the different transition processes,
the discretized observation states need to be divided into
multiple subsets according to the hourly average load mea-
surements. Each subset is used to train a Markov chain
model. Specifically, first, the entire collection of discretized
observation states is split into \( N_d \) subsets according to the
corresponding low-resolution load observation, \( P_d(t) \). The jth
subset is obtained as:

\[ D_j = \{S_t(i)\}, \quad i \in \{1, \ldots, N\}', \quad t \in \{1, \ldots, N\}, \]

where \( j \) is a subset index and \( \{ \cdot \} \) is a set.

\[ \text{if} \quad F\left( \frac{(j - 1) \times 100}{N_d} \right) \leq P_d(t) < F\left( \frac{j \times 100}{N_d} \right), \]

where, \( F(\cdot) \) is a function that returns percentiles of the entire
set of low-resolution load observations, and \( N' \) is the total number of discretized observation states in each low-resolution
sampling interval.

Then, for each subset \( D_j \), the stochastic process is parameter-
erized by empirically estimating the transition probabilities
between discrete observed states in terms of a transition
matrix. A second-order Markov process consists of three
states: the previous state, the current state, and the next state.
Therefore, the stochastic transition matrix, \( P_r \), is a three-
dimensional (3D) array, as illustrated in Fig. 6. Each element
of \( P_r \), \( P_r(x, y, z) \), represents the probability of moving to state
\( z \) under the condition that the previous state is \( x \) and the current
state is \( y \). For each subset \( D_j \), elements of \( P_r \) can be estimated
from the frequencies of posterior states. Assume \( D_j \) takes on
the form of \( \{S(i)\}, i = 1, \ldots, N' \), where \( N' \) is the total number
of observation states in \( D_j \), then the occurrence number at
\( (x, y, z) \) can be counted as:

\[ \sum_{i=2}^{N' - 1} \{S(i - 1) \Rightarrow x \} \]
\[ S(i) \Rightarrow y \quad \text{and} \quad S(i + 1) \Rightarrow z \],

where, \([\cdot]\) is the Iverson bracket which converts any logical
operation into 1, if the operation is satisfied, and 0 otherwise.
“\( \Rightarrow \)" stands for the “equal to” operator and “and” is the logi-
cal and operator. Thus, the elements of transition probability
matrix are computed by:

\[ P_r(x, y, z) = \frac{n(x, y, z)}{\sum_z n(x, y, z)}, \quad x, y = 1, \ldots, N'. \]

For each subset \( D_j, j = 1, \ldots, N_d \), (9) is performed to obtain
a 3D stochastic transition matrix. Moreover, for each service
transformer with a micro-PMU, the entire above-mentioned
procedure for parameterizing variability is conducted to obtain
\( N_d \) stochastic transition matrices.

III. ENRICHING LOAD DATA FOR TRANSFORMERS WITH
ONLY SMART METERS

A. Determining Teacher Weights

Recall that our goal is to recover the high-resolution load
data masked by the low-resolution load data. In this procedure,
we leverage teacher models of transformers with micro-PMUs
for service transformer with only SMs. Note that there might
be more than one teacher transformer serving the same number of customers as the student transformer supplies. Different teacher transformers have different load behaviors. Thus, it is necessary to determine the learning weights corresponding to particular teacher transformers. These weights are determined by evaluating customer-level load similarity between the teacher and student transformers using low-resolution load data.

Specifically, for the \( t \)th customer served by a student transformer, we can obtain a typical daily load pattern, \( P_i \), which reflects customer behavior and the total capacity of appliances [16]. Then, for a student transformer serving \( N_c \) customers, we can obtain \( N_c \) daily load patterns, \( \{P_1, \ldots, P_{N_c}\} \). Similarly, for a teacher transformer supplying \( N_t \) customers, we can obtain \( N_t \) daily load patterns. Since we have multiple teacher transformers, we can obtain a load pattern collection for the \( k \)th teacher transformer is denoted by \( \{P_{k1}, \ldots, P_{kN_t}\}, k = 1, \ldots, N_t \), where, \( N_t \) is the total number of teacher transformers. Then, load similarity between a student transformer and the \( k \)th teacher transformer is evaluated as:

\[
W_k = \frac{N_c}{\sum_{i=1}^{N_c} \sum_{j=1}^{N_i} |P_i - P_j|}, \quad k = 1, \ldots, N_t, \tag{10}
\]

where, \( N_i \) denotes the number of customers served by the \( k \)th teacher transformer. Thus, the teacher and student transformers do not necessarily serve the same number of customers. The \( W_k \)'s in (10) are then normalized for a more convenient mathematical representation:

\[
W_k = \frac{W'_k}{\sum_{k=1}^{N_t} W'_k}, \tag{11}
\]

In summary, the normalized similarity weights reflect the confidence of a student transformer to learn from multiple teacher transformers for load data enrichment.

B. Enriching Load Data

Using the normalized teacher weights, along with the load boundary and variability inference models derived in Section II, we can conduct poor load data enrichment for service transformers that only have SMs.

1) Inferring Load Boundaries: In Section II-A, for each transformer with high-resolution load data, we have obtained two GPR models for inferring the maximum and minimum loads given the corresponding average load over each low-resolution sampling interval. These two models are non-parametric and expressed in (4). Specifically, the trained maximum load inference model for the \( k \)th teacher transformer is expressed as:

\[
\begin{bmatrix}
\bar{P}_k(1) \\
\vdots \\
\bar{P}_k(N)
\end{bmatrix} = \begin{bmatrix}
f_k(P_{a,k}(1)) \\
\vdots \\
f_k(P_{a,k}(N))
\end{bmatrix} \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{12}
\]

To conduct load data enrichment, first, customer-level SM data are aggregated to obtain the load supplied by the student transformer, namely, \( \{P_{a,*}(1), \ldots, P_{a,*}(N)\} \). Note that the transformer loss is approximated and added to the aggregate load. Specifically, the total loss of a student transformer supplying an aggregate load, \( P_{a,*}(t) \), is estimated as follows:

\[
P_{l,*}(t) = P_{\text{full},*} + \frac{P_{a,*}(t)}{P_{\text{rate},*}} P_{\text{full},*}, \quad t = 1, \ldots, N, \tag{13}
\]

where, \( P_{\text{full},*} \) and \( P_{\text{full},*} \) denote the no-load loss and full-load loss, respectively. \( P_{\text{rate},*} \) denotes the kVA rating of the student transformer. \( P_{\text{full},*}, P_{\text{full},*} \), and \( P_{\text{rate},*} \) are typically provided by transformer manufacturers. Note that the effect of reactive power is ignored when estimating the loss because the reactive power is typically small [7]. For conciseness, we assume that \( P_{a,*}(t) \) in the following sections has already included the aggregate load and the corresponding total loss of the student transformer.

Then, we assume the unknown upper bound of instantaneous load in terms of a function variable, \( \bar{P}_{k,*}(t) = f_k(P_{a,*}(t)), t = 1, \ldots, N \), follows a Gaussian distribution. By appending \( \bar{P}_{k,*}(t) \) at the end of (12), an \( (N + 1) \)-dimensional Gaussian distribution can be formed as:

\[
\begin{bmatrix}
\bar{P}_k(1) \\
\vdots \\
\bar{P}_k(N) \\
\bar{P}_{k,*}(t)
\end{bmatrix} = \begin{bmatrix}
f_k(P_{a,k}(1)) \\
\vdots \\
f_k(P_{a,k}(N)) \\
f_k(P_{a,*}(t))
\end{bmatrix} \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{14}
\]

where, \( \Sigma_{k*} \) represents the training-test set covariances and \( \Sigma_{**} \) is the test set covariance. In (14), observations for \( \{\bar{P}_k(1), \ldots, \bar{P}_k(N)\} \) are known and denoted by \( \bar{P}_k = \)
Thus, using the Bayes rule, the distribution of \( \bar{P}_k(t) \) conditioned on \( \bar{P}_k \) is obtained as:

\[
\bar{P}_k(t | \bar{P}_k) \sim \mathcal{N}(\mu_k(t), \Sigma_k(t)),
\]

where, \( \mu_k(t) = \Sigma_k^T \Sigma_k^{-1} \bar{P}_k \) and \( \Sigma_k(t) = \Sigma_k - \Sigma_k^T \Sigma_k^{-1} \Sigma_k \).

Note that \( \mu_k(t) \) denotes the most probable value of the estimated upper bound of instantaneous load given the average load during each low-resolution sampling interval.

Since we have \( N_t \) teacher transformers, we can obtain a total of \( N_t \) estimated maximum load candidates, namely, \( \{ \mu_1^N(t), \ldots, \mu_N^N(t) \} \). Also, considering load similarity between the student transformer and teacher transformers, a weighted-averaging operation is performed on all the inferred maximum loads to calculate a final estimated upper bound of instantaneous load:

\[
\bar{P}_s(t) = \sum_{k=1}^{N_t} W_k \mu_k(t), \quad t = 1, \ldots, N.
\]

The same procedure introduced above is also applied to infer the unknown minimum load, \( \underline{P}_s(t) \), using the known average load over each low-resolution sampling interval. Once we have obtained the estimated load boundaries, then the trained probability matrices can be leveraged to infer load variability within each boundary.

2) Inferring Load Variability: As introduced in Section II-B, each teacher transformer has \( N_d \) extracted transition matrices corresponding to different load levels. Therefore, the first step in inferring the high-resolution load variability is to determine which transition matrix to use. In other words, we need to find the variability inference matrix corresponding to the load level that the low-resolution load measurement belongs to. This is achieved by splitting the known low-resolution load observations of the student transformer into \( N_d \) subsets:

\[
P_s = \{ P_{a,s}(t), \quad i \in \{1, \ldots, N\}, j = 1, \ldots, N_d, \}
\]

\[
\text{if } F \left( \frac{(j-1) \times 100}{N_d} \right) \leq P_{a,s}(t) < F \left( \frac{j \times 100}{N_d} \right).
\]

Then, the \( j \)th stochastic transition matrix of each teacher transformer is selected for enriching the low-resolution load measurements in the \( j \)th subset of the student transformer, \( P_s \). Moreover, considering that there is more than one teacher transformer, i.e., for each subset \( P_s \), we have \( N_t \) transition matrices to use. Thus, before proceeding to instantaneous load variability inference, a weighted averaging process similar to the load boundary estimation is conducted to obtain a comprehensive transition modal:

\[
P_{t,s}^j = \sum_{k=1}^{N_t} W_k P_{t,s}^{j,k}, \quad j = 1, \ldots, N_d.
\]

where, \( P_{t,s}^{j,k} \) stands for the transition matrix for the \( k \)th teacher transformer based on the \( j \)th subset of observation states, \( D_j \). Then, for each low-resolution load measurement to be enriched, \( P_{a,s}(t) \), the final targeted transition matrix, \( P_{t,s}^j \), and the inferred load boundary, \( (\bar{P}_s(t), \underline{P}_s(t)) \), are leveraged to generate the targeted high-resolution load data. Specifically, assume the previous state is \( S_{t,s}(i-1) \), and the current state is \( S_{t,s}(i) \), our goal is to determine the next state, \( S_{t,s}(i+1) \), where, \( i = 1, \ldots, N' \), stands for the sequence number of states within the \( t \)th low-resolution sampling interval. To do this, first, a random value at \( i, U_s(i) \), is generated from the uniform distribution within the interval \( (0, 1) \). Then, the state at \( (i + 1) \) is determined by:

\[
S_{t,s}(i + 1) = z_s, \quad i = 2, \ldots, N' - 1, \quad \text{if } \sum_{z=1}^{z_s-1} P_{t,s}^j S_{t,s}(i-1), S_{t,s}(i), z \leq U_s(i)
\]

\[
< \sum_{z=1}^{z_s} P_{t,s}^j (S_{t,s}(i-1), S_{t,s}(i), z).
\]

Note that the generated \( S_{t,s}(i) \)'s are discrete state samples, therefore, they need to be transformed to specific load samples:

\[
P_{t,s}^j(i) = \frac{P_t(i) + S_{t,s}(i) (\bar{P}_s(t) - \underline{P}_s(t))}{N_s}, \quad i = 1, \ldots, N'.
\]

Since there is more than one low-resolution time interval, the above procedure is conducted for each low-resolution load observation. Also, since the low-resolution load observations are grouped into multiple subsets, the entire procedure introduced above is conducted through all subsets of low-resolution load measurements. The detailed steps for load data enrichment for a student service transformer is illustrated in Fig. 7.

IV. CASE STUDY

In this section, we have validated the proposed load data enrichment approach using real high- and low-resolution load data [17].

A. Dataset Description

The dataset includes real 1-second load data for eight service transformers and hourly SM energy data for 185 customers. Among these customers, 36 are supplied by the 8 transformers with high-resolution load data (with micro-PMUs), and the remaining 149 customers are fed by the other 34 service transformers with low-resolution load data (with only SMs). To verify the performance of load data enrichment, the utility has also installed extra measuring devices to record 1-second load data for the service transformers with only SMs [17]. The time range of the dataset is two months. In practice, micro-PMUs might have higher sampling rates than one sample per second, however, there is no fundamental difference in verifying the performance of our approach.

B. Enriching Low-Resolution Load Measurements

Fig. 8 shows one-day actual and enriched 1-second load data for a service transformer. As can be seen, the enriched curve can accurately follow the actual basic load pattern. Note that our goal is not to force the enriched 1-second data to
It is also of interest to examine the results obtained using our proposed load data enrichment framework with a first-order Markov chain model. Fig. 12 presents the actual high-resolution load curve and the enriched load curve based on a first-order Markov model. To assess the different effects of the first- and second-order Markov models on load variation inference, we have constructed the distributions of load state duration, as shown in Fig. 13, where, $D$ denotes the load state duration. By comparing Figs. 13(b) and 13(c) with Table I, we observe that the computed maximum, minimum, median absolute errors of the percentiles are 2.7, 0.31, and 1.5, respectively. The RMSE is 1.6. Therefore, the differences are small, which also proves the effectiveness of our proposed approach from a statistical perspective.

Note that our final goal is to recover the statistical properties of the high-resolution load within each low-resolution sampling interval. Therefore, the performance of our proposed approach needs to be evaluated from the perspective of statistics. Fig. 10 illustrates the distributions of the actual and enriched load samples on the load curves shown in Fig. 8. It demonstrates that the enriched load distribution closely matches the actual load distribution. In comparison, the non-enriched load curve, which only includes 24 load observations, cannot sufficiently form a satisfactory distribution. In addition, to quantitatively assess load enrichment performance, we have examined the differences between the actual and enriched load values corresponding to different percentiles, as shown in Fig. 11. We have also evaluated the difference between the percentiles of the enriched load and the actual load. The computed maximum, minimum, median absolute errors of the percentiles are 2.7, 0.31, and 1.5, respectively. The RMSE is 1.6. Therefore, the differences are small, which also proves the effectiveness of our proposed approach from a statistical perspective.

### Table I - Computed Error Metrics of Inferring Load Boundary

<table>
<thead>
<tr>
<th></th>
<th>Maximum AE</th>
<th>Minimum AE</th>
<th>Median AE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>3.8</td>
<td>0.01</td>
<td>1.1</td>
<td>1.9</td>
</tr>
<tr>
<td>$P'$</td>
<td>2.7</td>
<td>0.06</td>
<td>1.1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Fig. 8. One-day actual and enriched 1-second load curves.

Fig. 9. The estimated maximum and minimum loads against the corresponding actual values.

Fig. 10. Distributions of the actual and enriched 1-second load in Fig. 8.

Fig. 11. Percentiles of the actual and enriched 1-second loads in Fig. 8.

Note that our final goal is to recover the statistical properties of the high-resolution load within each low-resolution sampling interval. Therefore, the performance of our proposed approach needs to be evaluated from the perspective of statistics. Fig. 10 illustrates the distributions of the actual and enriched load samples on the load curves shown in Fig. 8. It demonstrates that the enriched load distribution closely matches the actual load distribution. In comparison, the non-enriched load curve, which only includes 24 load observations, cannot sufficiently form a satisfactory distribution. In addition, to quantitatively assess load enrichment performance, we have examined the differences between the actual and enriched load values corresponding to different percentiles, as shown in Fig. 11. We have also evaluated the difference between the percentiles of the enriched load and the actual load. The computed maximum, minimum, median absolute errors of the percentiles are 2.7, 0.31, and 1.5, respectively. The RMSE is 1.6. Therefore, the differences are small, which also proves the effectiveness of our proposed approach from a statistical perspective.

It is also of interest to examine the results obtained using our proposed load data enrichment framework with a first-order Markov chain model. Fig. 12 presents the actual high-resolution load curve and the enriched load curve based on a first-order Markov model. To assess the different effects of the first- and second-order Markov models on load variation inference, we have constructed the distributions of load state duration, as shown in Fig. 13, where, $D$ denotes the load state duration. By comparing Figs. 13(b) and 13(c) with...
Fig. 12. One-day actual and enriched 1-second load curves (1st-order Markov model).

Fig. 13. Distributions of load state duration corresponding to the actual load and the enriched loads.

Fig. 14. Percentiles of load state duration corresponding to the actual load and the enriched loads.

Fig. 15. Robustness of our proposed approach to small-scale PVs.

In modern distribution systems, PV integration is common for utilities. Therefore, it is necessary to test the performance of our load data enrichment approach under the condition of PV integration. Specifically, three scenarios are considered where in all scenarios, both the teacher transformer and the student transformer supply six customers. In the first scenario, three of the six customers supplied by the teacher transformer have installed PVs, and the ratio of the peak PV generation to the peak load of the teacher transformer is 44%. In the second scenario, only the student transformer supplies three PV-installed customers, and the ratio of the peak PV generation to the peak load of the student transformer is 32%. In the third scenario, the teacher and student transformers both have three PV-installed customers, and the ratios of the peak PV generation to the peak loads of the teacher and student transformers are 40% and 34%, respectively. The enrichment results corresponding to the three foregoing scenarios are shown in Fig. 15. It is demonstrated that the proposed approach can still achieve accurate high-resolution load data enrichment when the teacher and/or student transformers serve PV-installed customers. Quantitatively, for the first scenario, the maximum, minimum, and median absolute errors between the percentiles of the enriched load and the actual load are 1.91, 0.23, and 0.96, respectively. For the second scenario, the three computed error metrics are 3.24, 1.85, and 2.40, respectively. For the third scenario, the three computed error metrics are 3.24, 1.85, and 2.40, respectively. In summary, the error metrics demonstrate that our proposed load data enrichment framework can adapt to PV integration.

D. Performing Time-Series Power Flow Studies

To thoroughly examine the performance of our proposed approach, we have conducted time-series power flow studies by separately feeding the actual and enriched loads into a real distribution system [18]. The one-line topology of the real distribution system is shown in Fig. 16. Bus voltages obtained from power flow analysis, which are critical to distribution system operators, are used to evaluate our proposed approach. Specifically, we compare the distributions of bus voltages and voltage ramps obtained from power flow studies based on the actual and enriched high-resolution load data, respectively. The reason for assessing voltage ramp is that voltage ramp is significant for renewable energy integration [8]. The voltage ramp
E. Performance Comparison

It is of significance to compare our approach with other methods presented in previous works. We primarily focus on comparing our approach with an allocation-based methodology introduced in [8] and a noise-based technique presented in [12], which are two primary load data enrichment approaches in previous works.

1) Comparison With the Allocation-Based Method: The allocation-based method involves two steps. First, a low-resolution substation- or feeder-level load profile is scaled to obtain service transformer-level load profiles, according to transformer capacity or peak load. Then, the scaled low-resolution load profile is enriched using a variability library, which is constructed by applying the discrete wavelet transform algorithm to known high-resolution transformer-level load measurements. An alternative to scaling low-resolution load profile is to obtain a load pattern obtained by scaling known typical load profiles of other transformers, as presented in [8]. For conciseness, we refer to the techniques presented in [8] as the allocation-based method. The performances of our approach and the allocation-based approach are shown in Figs. 18(a) and 18(b), respectively, where the actual and enriched load curves on a certain day are presented. In Fig. 18(a), we can observe that the basic pattern of the enriched 1-second load can flexibly follow the actual load variation, despite load uncertainty. The superior performance of our approach results from two aspects, the fine spatial granularity of SM data and the design of load boundary inference process. In comparison, the allocation-based load enrichment approach fails to accurately track the basic load pattern, as demonstrated in Fig. 18(b).

The performance of the allocation-based method can also be evaluated by examining the $R^2$ values computed for the load bounds, as shown in Fig. 19. We can observe that the $R^2$ values are negative, which means that the estimated maximum/minimum bound offers a poor estimation of the variation of the actual maximum/minimum bound. The unsatisfying performance of the allocation-based approach can also be viewed by observing the two scatter plots in Fig. 19, where most scatters are located above the upper-right diagonal line, indicating an overestimation of the actual load bounds.

To further evaluate the performance of our approach and the benchmark methods, we have also computed the cumulative probability of the actual and enriched load presented in Fig. 18. The empirical cumulative distribution functions (ECDFs) are illustrated in Fig. 21, where, we can observe that the ECDF corresponding to our method is much closer to the ECDF of the actual load than the ECDF corresponding to the allocation-based method. To quantitatively assess the similarity between the two ECDFs, we have computed the two-sample Kolmogorov-Smirnov (KS) statistic for each method, using the following equation:

$$D = \sup_P |F_a(P) - F_e(P)|,$$

where, sup denotes the supremum of the set of distances. $F_a(P)$ denotes the ECDF of the actual high-resolution load, and $F_e(P)$ denotes the ECDF of the enriched load. Intuitively
Fig. 18. The actual high-resolution load curve and the enriched load curves.

(a) Actual curve and the enriched curve using our approach

(b) Actual curve and the enriched curve using an allocation-based approach

(c) Actual curve and the enriched curve using a noise-based approach

Fig. 19. The estimated maximum and minimum load bounds obtained from the allocation-based method against the corresponding actual values.

(a) Maximum

(b) Minimum

Fig. 20. The estimated maximum and minimum load bounds obtained from the noise-based method against the corresponding actual values.

Fig. 21. Cumulative probability distributions of the actual load and the enriched load in Fig. 18.

TABLE II
COMPUTED ERROR METRICS BASED ON LOAD CURVES IN FIG. 18

<table>
<thead>
<tr>
<th>Our Approach</th>
<th>Allocation-based</th>
<th>Noise-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>nMAE (%)</td>
<td>12.3</td>
<td>22.8</td>
</tr>
<tr>
<td>nRMSE (%)</td>
<td>16.4</td>
<td>28.1</td>
</tr>
</tbody>
</table>

computed $D$ value for the noise-based method is 0.32, which is greater than 0.14, as shown in Fig. 21. This demonstrates that our method has a better performance than the noise-based method in terms of the two-sample KS statistic.

To quantitatively compare the aforementioned three approaches, we have also computed the normalized mean absolute error (nMAE) and the normalized root mean square error (nRMSE) based on the load curves in Fig. 18. Specifically, nMAE and nRMSE are computed as follows:

\[
nMAE = \frac{\sum_{t=1}^{n_t} |P(t) - \hat{P}(t)|}{n_t \cdot P_{max}} \times 100\%,
\]

\[
nRMSE = \frac{\sqrt{\sum_{t=1}^{n_t}(P(t) - \hat{P}(t))^2}}{n_t \cdot P_{max}} \times 100\%,
\]

where, $n_t$ is the total number of samples in a day with a resolution of 1 second, i.e., 86400. $P(t)$ and $\hat{P}(t)$ denote the actual and estimated loads at time $t$, respectively. $P_{max}$ denotes the peak of the actual load. The computed error metrics are summarized into Table II. We can see that compared to the allocation- and noised-based methods, our approach has smaller errors.

V. CONCLUSION

This paper is devoted to temporally enriching low-resolution load data for service transformers that only have SMs, using
high-resolution load data from service transformers with micro-PMUs and SMs. The entire process includes two stages, determining the maximum and minimum load bounds using known low-resolution load measurements and trained regression models, and inferring load variability within load boundaries using trained probabilistic transition models. The regression and transition models are trained using high-resolution load data from service transformers with micro-PMUs. We have used real high-resolution load data to prove that our approach is able to accurately recover high-resolution load data masked by the average load measurements over low-resolution sampling intervals. The enriched high-resolution load data can significantly enhance utilities’ grid-edge observability and situational awareness of distribution systems. Our paper’s key findings are summarized as follows.

- The 1-second load within an hourly interval can be 40% times larger or smaller than the corresponding average load during the same hour interval. By performing power flow studies, we have found that using the hourly average load for conducting power flow analysis cannot accurately capture the actual condition of distribution systems. Therefore, performing low-resolution power flow studies might cause significant errors, especially for those distribution networks that have a weak grid connection and long line segments.
- The numerical experiments have verified that our proposed approach shows strong robustness and adaptability to PVs.
- The numerical experiments have also demonstrated that our approach can accurately recover statistical properties of the instantaneous load within each low-resolution sampling interval of SM. The power flow studies show that our approach can faithfully reflect distribution system’s actual voltage conditions from a statistical perspective.

REFERENCES


Fankun Bu (Graduate 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 pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From 2008 to 2010, he worked as a Commissioning Engineer with NARI Technology Company Ltd., Nanjing, China. From 2013 to 2017, he worked as an Electrical Engineer with State Grid Corporation of China, Nanjing, Jiangsu, China. His research interests include distribution system modeling, smart meter data analytics, renewable energy integration, and power system protection.

Kaveh Dehghanpour (Member, IEEE) received the B.Sc. and M.S. degrees in electrical and computer engineering from the University of Tehran in 2011 and 2013, respectively, and the Ph.D. degree in electrical engineering from Montana State University in 2017. His research interests include application of machine learning and data-driven techniques in power system monitoring and control.

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology in 2008 and 2013, respectively. He is a Principal Investigator for a multi-state projects funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. His research interests include optimization and data analytics in power distribution systems and microgrids. He was a recipient of the National Science Foundation CAREER Award, the IEEE Power and Energy Society (PES) Outstanding Young Engineer Award, College of Engineering’s Early Achievement in Research Award, and the Harpole-Pentair Young Faculty Award Endowment. He is the Chair of IEEE PES PSE01 Award Subcommittee, the Co-Chair of PES Distribution System Operation and Planning Subcommittee, and the Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Associate Editor of IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, IEEE POWER ENGINEERING LETTERS, and IET Smart Grid.