A Novel Power-Band Based Data Segmentation Method for Enhancing Meter Phase and Transformer-Meter Pairing Identification

Han Pyo Lee, Senior Member, IEEE, PJ Rehm, Matthew Makdad, Edmond Miller, Associate Member, IEEE, and Ning Lu, Fellow, IEEE

Abstract—This paper presents a novel power-band-based data segmentation (PBDS) method to enhance the identification of meter phase and transformer-meter pairing. Meters that share the same transformer or are on the same phase typically exhibit strongly correlated voltage profiles. However, under high power consumption, there can be significant voltage drops along the line connecting a customer to the distribution transformer. These voltage drops significantly decrease the correlations among meters on the same phase or supplied by the same transformer, resulting in high misidentification rates. To address this issue, we propose using power bands to select highly correlated voltage segments for computing correlations, rather than relying solely on correlations computed from the entire voltage waveforms. The algorithm’s performance is assessed by conducting tests using data gathered from 13 utility feeders. To ensure the credibility of the identification results, utility engineers conduct field verification for all 13 feeders. The verification results unequivocally demonstrate that the proposed algorithm surpasses existing methods in both accuracy and robustness.

Index Terms—Data segmentation, ensemble clustering, machine learning, phase identification, smart meter data analysis, topology identification, transformer identification.

I. INTRODUCTION

A CCURATE meter phase and meter-transformer pairing information is crucial for efficient distribution circuit operation, planning, and asset management in modern utility systems. Manual registration of this data is error-prone, and updates are frequently required for circuit upgrades, reconfiguration, and equipment replacements. However, the large number of distribution transformers and customer meters makes manual verification costly. As a result, utility engineers are actively exploring data-driven, automated methods for periodic information verification.

The increasing adoption of smart meters [1] has provided utilities with time-series power and voltage measurements at sampling rates of 15 or 30 minutes [2]. This availability of data has made leveraging smart meter data for automated identification of phase and transformer-meter pairing relationships an appealing option in recent years. Table I presents a comprehensive overview of existing methods for addressing phase and transformer-meter pairing identification problems, along with a comparison of their strengths and weaknesses in relation to the proposed approach introduced in this paper.

As shown in the Table I, in the realm of phase identification problems, there are two main approaches developed: directly measuring methods [3, 4] and data-driven methods. With the widespread deployment of smart meters, data-driven methods have become the preferred choice for utilities. Table I provides an overview of the various data-driven approaches, including real power-based [5, 6], voltage-based [7, 8], and machine learning-based approaches [9, 10, 11, 12, 13, 14].

The real power-based methods assume that the cumulative customer loads on a phase should align with the power measurements at the substation. Thus, a significant drawback to this approach is its dependency on 100% smart meter penetration. Additionally, since line losses are not factored into smart meter measurements, the misplacement of a small number of customers on a phase may remain undetectable. On the other hand, the voltage-based methods assume that customers on the same phase exhibit more pronounced correlations in voltage time-series profiles compared to those on different phases. Nevertheless, the current drawback of these voltage-based approaches lies in relying solely on voltage measurements for computing correlation, thereby overlooking the impact of differing customer power consumption levels on the degree of correlation in voltage profiles. Hence, in this paper, we propose to use power consumption levels to choose highly correlated voltage profile segments. This approach enables the utilization of both power and voltage data, contributing to enhanced identification accuracy.

Notably, machine learning (ML)-based methods have witnessed a surge in popularity in recent years. Building upon our prior work presented in [15], which introduced a machine-learning based phase identification method, this paper extends...
the methodology to use real smart meter data as inputs instead of synthetic data. Furthermore, we validate its performance by conducting field verification executed by utility engineers, thereby affirming its superiority over existing methods.

In the realm of transformer-meter pairing, there are three main approaches: real power-based, voltage-based, and linear regression-based techniques, also highlighted in Table I. The process of identifying transformer-meter pairings shares noticeable similarities with meter phase identification. Among the real power-based methods, the load-summing approach [6], [16] has gained significant popularity. Moreover, voltage correlation-based methods are also commonly used for transformer-meter pairing identification [17], [18], [19], [20]. The limitation common to both power- and voltage-based approaches is their failure to integrate voltage and power measurements to enhance identification accuracy. In recent years, linear regression-based methods have attracted increasing attention due to their mathematical simplicity in system topology identification [21], [22], [23]. However, this approach necessitates reactive power from AMI.

Upon comparing the methods, it becomes evident that the calculation of voltage correlation plays a pivotal role in meter topology identification, regardless of whether it pertains to phase identification or pairing identification. However, current state-of-the-art techniques suffer from two significant limitations. Firstly, all existing methods rely on utilizing the entire time-series data for voltage correlation computation. Nevertheless, as demonstrated in Section II, not all segments of the voltage profile exhibit high correlation among meters sharing the same phase or supplied by the same transformer. Incorporating numerous less-correlated voltage segments leads to a deterioration in correlation, resulting in increased false positive rates. Secondly, the majority of algorithms focus solely on voltage measurements as inputs, disregarding the valuable information provided by power measurements. Leveraging this additional data could effectively...

---

**TABLE I**

A REVIEW OF EXISTING METHODS AND OUR CONTRIBUTIONS

| Category                     | Methodology                        | Description                                                                 | Strength                                        | Weakness                                      |
|------------------------------|------------------------------------|-----------------------------------------------------------------------------|-------------------------------------------------|-----------------------------------------------|
| **Transformer-Meter Pairing** | **Identification**                 |                                                                             |                                                 |                                               |
| **Phase Identification**     | **Real power-based**               | Customer loads on the same phase should match substation power measurements | Simplicity                                      | Degradation depending on the AMI penetration |
|                              | **Voltage-based**                  | Customers on the same phase exhibit stronger voltage correlations than those on different phases | Easily-interpretable                            | Inefficient use of data                       |
|                              | **ML-based**                       | Unsupervised learning algorithms that solve the clustering problem          | High accuracy                                   | Uninterpretable                               |
|                              | **PBDS (Proposed)**               | Power-band based data segmentation algorithm to extract highly correlated voltage segments on the same phase | Easily-interpretable                            | Unexplainable                                 |
|                              | **Load summing**                  | Customer loads under the same DT should match DT power measurements       | Simplicity                                      | DT metering required                          |
|                              | **Voltage-based**                 | Analyze voltage correlation coefficients between all customers labeled on the same transformer | Easily-interpretable                            | Inefficient use of data                       |
|                              | **Regression-based**              | Estimate line impedance using linear regression and build up a secondary circuit | Mathematical simplicity                        | No solution provided for incorrect DT labeling |
|                              | **PBDS (Proposed)**               | A two-stage algorithm based on PBDS to distinguish between a group of meters under the same DT and a group of meters under different DTs | Present a method to correct abnormal DT-meter pairs | Requires reactive power from AMI              |
enhance the accuracy and robustness of voltage-based identification algorithms.

Therefore, in this paper, we present a novel power-band-based data segmentation (PBDS) method that leverages both power and voltage inputs. To establish the foundation for our proposed algorithm, we conduct circuit analysis to illustrate the occurrence of voltage correlation deterioration phenomena and to provide a theoretical framework for the application of the PBDS method. This new approach using power-bands and minimum duration thresholds to select highly correlated voltage data segments from the time-series voltage profiles of a pair of meters to obtain an ensemble of segments. Then, voltage correlations of the two meters can be calculated from the ensemble instead of using two entire voltage profiles.

The primary contribution of the PBDS approach is the use of power-bands for data segmentation, leading to a substantial improvement in the accuracy of downstream tasks such as clustering phase groups or identifying transformer-meter pairs. The concept of selecting highly correlated voltage segments using power bands, along with a duration threshold for excluding short data segment, is a novel idea that has not yet been explored in existing literature. To evaluate the effectiveness of this approach, extensive testing was conducted on 13 real utility feeders. Most importantly, the performance improvement achieved by our approach, compared to existing methods, has been fully validated through extensive field verification conducted by utility engineers.

The rest of this paper is organized as follows. Section II introduces the phase and transformer-meter pairing identification problems, the correlation deterioration phenomena, and PBDS principles. Section III presents the application of PBDS in meter phase identification and in transformer-meter pairing identification. Simulation results are presented in Section IV and Section V concludes the paper.

II. METHODOLOGY

In this section, we provide an introduction to two fundamental problems in meter topology identification: meter phase and transformer-meter pairing. We then proceed to discuss the voltage correlation deterioration phenomena and elaborate on the utilization of PBDS as a means to mitigate this issue. Lastly, we present the PBDS based meter phase and transformer-meter pairing algorithms.

A. Meter Topology Identification Problems

1) Problems Description: In contemporary power distribution systems, the prevalence of two-phase loads has significantly diminished, primarily found in older residential installations and less common in new construction projects [24]. Therefore, this paper concentrates exclusively on the identification of phases for 1-phase loads. Specifically, our investigation focuses on single-phase, two-winding transformers operating within a 60 Hz AC system with a line-to-ground voltage of 7.2 kV. This configuration is representative of the prevalent setup in the power distribution systems of the United States. These transformers power homes at 240 V in a 120/240 V split-phase system.

As illustrated in Fig. 1, in a power distribution system serving multiple 1-phase circuits, the meter topology has two crucial pieces of information: the phase of the circuit and the distribution transformer supplying power to the meter. The phase information holds vital importance for utility engineers as it facilitates the assessment of 3-phase load balancing conditions across a distribution circuit. Similarly, transformer-meter pairing information enables utility engineers to estimate the transformer loading conditions.

2) Data Requirement: The input data for the phase identification algorithm using voltage correlation-based approaches are presented in Table II. Typically, voltage measurements taken at customer meters serve as the primary inputs for phase identification algorithms, as demonstrated in previous methods such as [12], [13], [14]. Additionally, in [9], [10], voltage measurements at the feeder head can be incorporated to enhance the voltage correlation calculation. Conversely, for transformer-meter pairing identification, the input data typically consists of only smart meter data and the corresponding transformer labels assigned to each smart meter.

In comparison to existing methods, our approach achieves significant performance improvements by incorporating both power and voltage data. It is important to note that both types of data are readily available to utilities without the need for additional metering. This approach allows for the identification of voltage segments that display stronger correlations among meters sharing the same phase or supplied by the same transformer. Not utilizing both voltage and power measurements at the meter level not only results in an inefficient use of data but
also contributes to a high false positive rate, as demonstrated in the results section.

3) Voltage Correlation Calculation: The underlying assumption of the voltage correlation based approach is that, when two 1-phase customers (indexed by \( i \) and \( j \)) are on the same phase or supplied by the same 1-phase DT, the correlation between their voltage profiles, \( V^i \) and \( V^j \), are stronger than otherwise would have been.

As will be explained in Section II-C, we calculate correlations between many short voltage segments (each within the range of a few hours) on two voltage profiles in order to determine the overall correlation between the two curves. This makes Pearson Correlation Coefficient (PCC) sufficient for computing correlations instead of using sophisticated non-linear correlation methods. PCC between \( V^i \) and \( V^j \), is calculated as

\[
PCC(V^i_n, V^j_n) = \frac{\sum_{n=1}^{N}(V^i_n - \bar{V}^i)(V^j_n - \bar{V}^j)}{\sqrt{\sum_{n=1}^{N}(V^i_n - \bar{V}^i)^2 \sum_{n=1}^{N}(V^j_n - \bar{V}^j)^2}}
\]

where \( \bar{V} \) is the average voltage, \( n \) is the \( n^{th} \) data point, and \( N \) is the number of points on the waveform.

Once the PCCs between meter pairs are calculated, we can either use a wide variety of classification methods to group meters together for phase group identification or compare the pair-wise PCCs to determine transformer-meter pairs.

B. Voltage Correlation Deterioration Phenomenon

1) Circuit Analysis: As shown in Fig. 2(a), in relation to each other, two 1-phase loads connected to the same 1-phase transformer have three basic connection types: in-parallel, partially-parallel, and in-series. Denote \( V_T \), \( I_T \), and \( R \) as the DT voltage, shared line current and resistance, respectively. Denote \( V_i \) and \( V_j \), \( I_i \) and \( I_j \) as the system voltage and current of loads \( i \) and \( j \), respectively. Denote \( R_i \) and \( R_j \) as the resistance of the transformer secondary connection to loads \( i \) and \( j \).

Then, we have

\[
V_i = V_T - IR - I_iR_i \quad (2)
\]

\[
V_j = V_T - IR - I_jR_j \quad (3)
\]

\[
I = I_i + I_j \quad (4)
\]

Note that for type 1, \( R = 0 \); for type 3, \( R_i = 0 \).

From (2)-(4), we have the following insights.

- When the transformer voltage is maintained close to its nominal values, the current on the secondary circuits (i.e., the lines connecting the transformer to the customer meters), \( I_i \) and \( I_j \), are mainly determined by the power consumption levels of the \( i^{th} \) and \( j^{th} \) loads, \( P_i \) and \( P_j \). Therefore, if \( P_i \) and \( P_j \) are both low, then \( I_i \) and \( I_j \) are low, causing a very small secondary voltage drop. Then, \( V_i \) and \( V_j \) will be very close to \( V_T \), making \( V_i \) and \( V_j \) highly correlated.
- However, when \( P_i \) and \( P_j \) are increasing, \( I_i \) and \( I_j \) will increase, leading to high secondary voltage drops on the secondary circuits. This will weaken the correlation between \( V_i \) and \( V_j \) significantly because the voltage variations are mainly determined by the values of \( R, R_i, R_j \) and \( |P_i - P_j| \).

To illustrate the phenomenon of voltage correlation deterioration with respect to the increase of local load consumption, Monte Carlo simulations are conducted to calculate PCCs between \( V_i \) and \( V_j \) for the three basic load connection types when the power levels of loads \( i \) and \( j \) vary randomly between 0 and 16 kW; \( V_T \) varies randomly from 122 V within five bands (i.e., 0.2, 0.4, 0.6, 0.8, and 1 V); and \( R, R_i \) are the service drop resistance provided by the utility; \( R_j \) are set to 0.01 \( \Omega \), 0.05 \( \Omega \), and 0.05 \( \Omega \), respectively.

As shown in Fig. 2(b), the voltage correlation deterioration for type 1 connection is the most prominent. When the output levels of both loads are above 6 kW while the variation of \( V_T \) is small (e.g., 0.2 V against 122 V), the correlation will drop below
0.4. The voltage correlation deterioration can exacerbate when the two loads are served by different DTs on the same phase due to the increase in line impedance. Therefore, excluding data sets that cause correlation deterioration is crucial for improving the accuracy of voltage-correlation based phase identification methods.

Next, we analyze the distribution of PCCs between loads \( i \) and \( j \) for low and high power bands under different transformer-level voltage fluctuations. Note that many 1-phase loads are residential. At night or daytime, when occupants are sleeping or not at home, the total electricity consumption is within a band of [0 2] kW. For a load with no data segment below 2 kW, we can widen the band. As shown in Fig. 2(c), for voltage segments between low power bands (e.g., 0–2 or 2–4 kW), the secondary voltage drops are very small when \( P_i \) and \( P_j \) are low. This makes \( V_i \) and \( V_j \) nearly equal to \( V_T \). As a result, the correlation between \( V_i \) and \( V_j \) is the strongest.

However, when we increase the power band from 2 kW to 16 kW at a 2-kW step, the PCC value deteriorates, especially for type 1 connection. This is because, for loads with high electricity consumption levels, the secondary circuit voltage variations increase to a level equal to or even larger than the transformer-level voltage variations, making voltage correlations between loads with different power consumption patterns weaker. Therefore, we conclude that, for meter phase identification, selecting voltage segments within a low power band will increase identification accuracy by mitigating voltage correlation deterioration.

### 2) Voltage Statistics From Actual Smart Meter Data

To determine the topology relationship of a pair of meters, we need to know if they are served by the same 1-phase transformer (\( M_{0-M_1} \)), by different transformers on the same phase (\( M_{0-M_2} \)), or by different transformers on different phases (\( M_{0-M_3} \)). This shows for meter phase identification, using a low power band is preferred.

Using a wide power band (\( P > 2 \) kW), the average PCC of \( M_{0-M_1} \) shows a greater separation from the average PCCs of \( M_{0-M_2} \) and \( M_{0-M_3} \). This disparity arises due to the fact that, for high power consuming loads, the voltage fluctuations between loads supplied by different DTs are more significant compared to those between loads served by the same DT. Consequently, the voltage correlation becomes weaker. These findings highlight the preference for a wider power band when aiming to identify transformer-meter pairings. The validity of this observation is further supported by the results summarized in Table III.

#### C. Power-Band Based Data Segmentation (PBDS) Algorithm

In [7], [8], the authors have proven that using voltage correlation for clustering is an effective method for customer phase identification. Based on the insight gained from the circuit analysis in Section II-B, we further demonstrate that using PBDS algorithm for extracting data segments can obtain data segments with stronger correlation between two time series \( V_i \) and \( V_j \).

Define a power band as \([P^- P^+]\), where \( P^- \) and \( P^+ \) are the power low and high limits, respectively. To ensure that the segments are sufficiently long to compute the PCC between two time series data sets, the minimum length, \( T_{dur} \), is used for excluding short data segments. Note that if less than two segments are found, the entire time series will be utilized for computing the voltage correlations. Thus, the process for selecting data segments can be formulated as

\[
P^- \leq P_{i,t} \leq P^+, \quad T_{dur} \leq m_{i,k} \Delta T
\]

\[
P^- \leq P_{j,t} \leq P^+, \quad T_{dur} \leq m_{j,k} \Delta T
\]

\[
PCC(V_i^M, V_j^M) = \frac{\sum_{k=1}^{K} (V_{i,k} - \bar{V}_i)^2 (V_{j,k} - \bar{V}_j)^2}{\sqrt{\sum_{k=1}^{K} (V_{i,k} - \bar{V}_i)^4} \sqrt{\sum_{k=1}^{K} (V_{j,k} - \bar{V}_j)^4}}
\]

\[
D(V_i^M, V_j^M) = 1 - |PCC(V_i^M, V_j^M)|
\]
TABLE IV

PCC STATISTICS OF DATA SEGMENTS BY THE \([P^{-}P^{+}]\) AND \(T_{dur}\) FOR FEEDER 10

| \(T_{dur} [h]\) | Power band \([P^{-}P^{+}]\) | \(\mu\) | \(\sigma\) | \(\%n\) | \(\mu\) | \(\sigma\) | \(\%n\) | \(\mu\) | \(\sigma\) | \(\%n\) | \(\mu\) | \(\sigma\) | \(\%n\) |
|------------|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.5        | 0~2 [kW]      | 0.48 | 0.10 | 38.4 | 0.29 | 0.25 | 16.2 | 0.27 | 0.18 | 34.2 | 0.27 | 0.18 | 34.2 |
| 1.0        | 0~4 [kW]      | 0.48 | 0.09 | 21.4 | 0.29 | 0.29 | 14.1 | 0.28 | 0.19 | 15.4 | 0.18 | 0.18 | 16.0 |
| 1.5        | 2~8 [kW]      | 0.49 | 0.09 | 14.4 | 0.46 | 0.25 | 19.3 | 0.34 | 0.20 | 13.0 | 0.33 | 0.20 | 13.1 |
| 2.0        | 2~10 [kW]     | 0.51 | 0.09 | 10.7 | 0.51 | 0.23 | 18.8 | 0.39 | 0.21 | 13.1 | 0.38 | 0.20 | 12.8 |
| 2.5        | 2~15 [kW]     | 0.54 | 0.09 | 8.8  | 0.53 | 0.21 | 16.9 | 0.43 | 0.20 | 12.6 | 0.42 | 0.20 | 12.4 |
| 3.0        | 2~20 [kW]     | 0.56 | 0.09 | 6.7  | 0.54 | 0.21 | 14.7 | 0.46 | 0.20 | 11.7 | 0.44 | 0.20 | 11.5 |

where \(i\) and \(j\) are the indices of meter \((i, j \in \{1, \ldots, N_M\})\), \(t\) is the time steps \((t \in \{1, \ldots, T\})\), \(\Delta T\) is the sampling interval, \(\bar{V}_i^M\) and \(\bar{V}_j^M\) are the mean values of \(V_i^M\) and \(V_j^M\), \(k\) is the index of the data segment \((k \in \{1, \ldots, K\})\), \(m_k\) is the number of data points in the \(k\)th data segments, \(M = \{m_1, \ldots, m_K\}\) is the set of the data segments. (7) computes the PCC matrix; (8) computes the correlation distance.

Note that (5) and (6) are the PBDS criteria (i.e., upper and lower power thresholds, and minimum duration). The data segment selection using low-power band \((0 \leq 2\ kW)\) and various \(T_{dur}\) is shown in Fig. 4. Only line segments between \([P^{-}P^{+}]\) with length longer than \(T_{dur}\) are selected. Table IV shows the PCC statistics of data segments that satisfy \([P^{-}P^{+}]\) and \(T_{dur}\). \(\mu\), \(\sigma\), and \(\%n\) denote the mean, standard deviation, and share of total segments, respectively. It is essential to select an optimal parameter that reflects the characteristics of the feeder. The selection of the best \([P^{-}P^{+}]\) and \(T_{dur}\) values for phase identification and transformer-meter pairing will be discussed in Section III.

As shown in Fig. 5, when applying PBDS, PCCs calculated between meter pairs on the same phase show much stronger correlations than those on different phases, i.e., Fig. 5(b) showing a much better contrast than Fig. 5(a).

III. METER TOPOLOGY IDENTIFICATION ALGORITHMS

In this section, we will present two PBDS based meter topology identification algorithms: phase identification and transformer-meter pairing.

A. Phase Identification

The key steps of the segmentation-based phase identification algorithm have been presented in our previous paper [15]. As shown in Fig. 6, the inputs of the algorithms are smart meter measurements including 15-min real power and voltage...
measurements. The PBDS steps are highlighted in the two shaded boxes. In this algorithm, we used the lowest power band, [0 2] kW, for extracting the voltage segments. This is because the correlation between meters on the same phase is the strongest within this band, as illustrated in Figs. 2 and 3. Then, the selected data segments are used for computing PCC and correlation distance between each pair of nodes. Based on correlation distances, the hierarchical clustering method will divide the smart meters into $3 \times n$ clusters with $n$ increasing from 1 to $N$ (i.e., the number of clusters increases from 3 to $3N$).

In this paper, we opted for hierarchical clustering over k-means clustering to mitigate the dependence on initial values, a limitation associated with k-means-based methods [25]. Additionally, given the radial structure of the power distribution system, hierarchical clustering facilitates the natural grouping of meters based on topology connections. This method is beneficial for achieving robust clustering results regardless of the initial placement of data points, especially in scenarios where phase labels are unknown.

If the phase labels are known, the phase of each 1-phase distribution transformer is provided by the utility, and the phase label on each meter is the same as that of the transformer to which it is connected. The majority vote mechanism presented in [9] will be used to assign phase labels to each cluster. The phase label of each meter is the same as the most dominant phase label in the cluster to which it belongs. And then, the prediction accuracy is evaluated based on the percentage of customers assigned the correct labels compared to the phase labels recorded by the utility. If the phase labels are unknown, the objective of the phase identification problem shifts to grouping customers who are most likely on the same phase together in a cluster for a given number of clusters. The CTS matrix constructed by the clustering ensemble method is proposed to determine the final clusters for utility engineers to label the phase for each cluster rather than for each individual customer, thereby reducing the number of field inspections from hundreds or thousands to tens of sites. Evaluating prediction accuracy is the same as in the previous case. Please refer to [15], [26] for more details.

### B. Transformer-Meter Pairing Identification

When there are multiple meters connected to the same transformer, a meter may be strongly correlated with a couple of meters but not with all of them. Therefore, we can build a map based on the voltage PCC values to estimate their connections. As shown in Fig. 7, $M_{11} \sim M_{14}$ belong to $T_1$ and $M_{21} \sim M_{23}$ belong to $T_2$. A connectivity map can be plotted by connecting the strongly correlated meters sequentially, starting with the meter with the maximum voltage magnitude (e.g., $M_{11}$, $M_{21}$) under the same transformer. If $M_{14}$ has a weaker correlation with other meters under $T_1$ than with meters under $T_2$, $M_{14}$ and the two transformers will be flagged as having abnormal connections.

As shown in Algorithm 1, 15-minute smart meter data (including power and voltage measurements) are used to compute the PCC between meters for the same 1-phase transformer and with all other 1-phase transformers for each season in a year. Specifically, we examine transformers serving multiple customers, i.e., those connected to more than one meter. For the same transformer, based on PCC, a transformer-meter connection map can then be drawn.

The transformer-meter pairing algorithm is a 2-stage process. In the first stage, the average PCC of a meter with respect to all the other meters under the same transformer is compared with the average PCC of the meter with all meters under another transformer to determine if the meter voltage is strongly correlated with its own meter group or another meter group. Because meters under the same transformer are not all strongly correlated, the average PCC of the top 2 correlated meters under the same transformer is compared with the average PCC of the top 2 meters under another transformer to identify mislabels that are not detected by the average PCC method. In both methods, if a transformer has a meter that is more strongly correlated with meters under another transformer, both transformers and their meters will be flagged for further inspection. Next, a seasonal PCC check is applied to identify moved meters. In the second stage, PBDS method is used to cross check to verify whether the meter is indeed mislabeled.

After this, the location of all mislabelled meters will be identified on the google-map based geographical information system to verify their locations by utility engineers to determine if a field check is necessary. We will explain the flagging and pairing verification algorithm in the following subsections.

#### 1) Flag Abnormal Transformer-Meter Pairs Using PCC Averages:

The average PCC for meter $j$ under transformer $i$ to all other meters under transformer $i$, $APCC_{i-j}$, can be computed as

$$APCC_{i-j}(j) = \frac{1}{N_i - 1} \sum_{k=1, k\neq j}^{N_i} PCC(V_j, V_k)$$

where $N_i$ is the number of meters supplied by transformer $i$.

The average PCC for meter $j$ under transformer $i$ to all other meters under transformer $r$, $APCC_{i-r}$, can be computed as

$$APCC_{i-r}(j) = \frac{1}{N_r} \sum_{k=1}^{N_r} PCC(V_j, V_k)$$

where $N_r$ is the number of meters supplied by transformer $r$. 

---

| PCC | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{14}$ |
|-----|---------|---------|---------|---------|
| $M_{11}$ | 1       | 0.96    | 0.93    | 0.84    |
| $M_{12}$ | -       | 1       | 0.95    | 0.84    |
| $M_{13}$ | -       | -       | 1       | 0.81    |
| $M_{14}$ | -       | -       | -       | 1       |

(a) 

![Fig. 7. Detected error connection on Feeder 6. (a) PCC Matrix of transformers 1 and 2. (b) Connectivity map based on PCC.](image-url)
Thus, if $APCC_{i-j} < APCC_{i-j}$, we will flag transformer $r$ as a possible host transformer for meter $j$ under transformer $i$. The IDs of the flagged meter, its current pairing transformer, the identified, possible pairing transformer will then be stored into three index vectors, $ID_{meter}$, $ID_{originalT}$, and $ID_{identifiedT}$.

If there are more than one transformer satisfies the condition, the one with the biggest $APCC_{i-j}$ value will be selected. In our study, we observed that not all meters served by the same transformer are strongly correlated. Thus, using $APCC_{i-j} < APCC_{i-j}$ as the only criteria may skew the results if one or two meters are not strongly correlated with the other meters.

2) **Flag Abnormal Transformer-Meter Pairs Using the Average of top-2 PCCs:** To increase the identification accuracy, the weighted average of the top-2 PCC values, $T2PCC$, is calculated for meter $j$ of transformer $i$ using the PCCs from the top 2 correlated meters under transformer $i$ and under transformer $r$, respectively, as

$$T2PCC_{i-j} = \sum_{k=1}^{2} w_k PCC(V_j, V_k)$$  \hspace{1cm} (11)

where $w$ is the weighting factor. Thus, we can use $T2PCC_{i-j} < T2PCC_{i-j}$ to flag transformer $r$ as a possible host transformer for meter $j$ under transformer $i$. Similarly, the IDs of the meter, its current pairing transformer, the identified pairing transformer will be stored into index vectors, $ID_{meter}$, $ID_{originalT}$, and $ID_{identifiedT}$ for further analysis.

3) **Flag Abnormal Transformer-Meter Pairs Using Seasonal PCC Variations:** Often times, the mislabelled transformer-meter pairing happens because a meter still in good condition has been moved from one transformer to another without its record being updated. Thus, when using the yearly data to compute $APCC$ or $T2PCC$, the values are much lower than what it should have been. This is because the PCC of the meter shows strong correlation with other meters under the labelled transformer for a period of time in a year, but after being moved away to another transformer, the correlation can drop drastically. Such transformer-meter pairing issues can be identified either by analyzing the time-series characteristic of $APCC$ or $T2PCC$ or by comparing the seasonal changes of $APCC$ or $T2PCC$.

In this paper, to identify meters with lower than average PCC values, we compute the seasonal average PCC of all meters served by transformer $i$, $APCC_i$, as

$$APCC_i = \frac{1}{N_i} \times \left( (N_i - 1) \sum_{j=1}^{N_i} \sum_{k=1,k\neq j}^{N_i} PCC(V_j, V_k) \right)$$  \hspace{1cm} (13)

Note that the seasonal PCC threshold ($\epsilon_{th}^{szn}$) is selected to detect sudden correlation drops. In this paper, based on the PCC distribution shown in Fig. 8, we set $\epsilon_{th}^{szn}$ to be 20th percentile so that all meters with $APCC_i < \epsilon_{th}^{szn}$ are flagged. In practice, $\epsilon_{th}^{szn}$ can be selected a little higher or lower depending on the cost of field verification.

4) **Identify the Possible Transformer(s) Supplying the Flagged Meter:** It is possible that for a mislabelled transformer-meter pair, we identify more than one transformer that may supply the possibly mislabelled meter, especially if we are using different criterion listed above. However, we find that
it is sufficient to select the most likely transformers to be the one with the highest APCC and T2PCC values calculated by (9) and (11), and then confirm the selection by using seasonal APCC. In theory, we can conduct exhaustive search for all transformers that satisfy APCC_{i-r}(j) < APCC_{i-r}(j) or T2PCC_{i-r}(j) < T2PCC_{i-r}(j). However, in practice, this approach is often times unnecessary because the 2-stage checking will further verify the mislabelled pair and the subsequent field inspection can quickly reveal the actual transformer-meter pairing information. Thus, it is crucial to identify the mislabelled transformer-meter pairs and less crucial to use the algorithm for identifying what transformer actually supplies the mislabelled meter.

5) PBDS Based Second Stage Verification: The goal of the 2-stage verification is to eliminate false positives. As shown by the correlation analysis in Section II-B, a meter with very lower power consumption tends to have high correlation with another meter on the same phase even though the two meters are supplied by different transformers. Because in such cases, the local power variations are very small, which cannot lead to sufficiently large APCC gaps.

Therefore, in the second stage verification process, we first exclude meters with very low power consumption when computing for APCC and T2PCC using (9), (10), (11), and (12). This can be achieved by using the data segments that satisfy 

\[ P \geq P_{TH} \]

where \( P_{TH} = 1 \) kW is the low power threshold. Again, to guarantee the length of the data segment is sufficient for computing PCC, we set the \( T_{dur} \geq 1 \) h.

Thus, for a meter, if there is no data segment that can satisfy \( P > 1 \) kW, it will be marked as a “low consumption meter”. For a flagged meters, if there is no data segment satisfies this condition, we will unflag the meter. If after the low power segments are removed, for a flagged meter, we have APCC_{i-r}(j) > APCC_{i-r}(j) or T2PCC_{i-r}(j) > T2PCC_{i-r}(j), the meter and the corresponding transformer will be unflagged. The remaining transformer-meter pairs will go through a GIS check and subsequent field verification.

IV. Simulation Results

One year of 15-minute smart meter data collected from 3,961 customers on 13 real feeders in the North Carolina area are used to test the performance of the two developed algorithms. The utility also provided the phase label for each 1-phase transformer. Within this dataset, voltage segments that satisfy the criteria of a power band ranging between \([0 \ 2]\) kW and a minimum duration of \(1\) h or more were identified in 99.9% of cases. If there were less than two voltage segments in the ensemble, the full time-series voltage profiles will be used to calculate the voltage correlations between the two meters. All tests were carried out on a desktop computer equipped with an Intel Core i7-7700@3.6 GHz processor and 16 GB of memory.

A. Case 1: Phase Identification With Phase Labels

To select the optimal parameters for each feeder, we conduct Monte Carlo simulations for the range of parameters shown in

![Fig. 9. Phase identification accuracy for three different numbers of clusters with varying parameters in Feeder 3. The accuracy of most parameter combinations converged when 3n∗ = 6. (a) 3 Clusters. (b) 6 Clusters. (c) 9 Clusters.](image)

Table V. We vary \( P^+ \) within the range of \([0 \ 2]\) kW with a 0.1 kW increment, keeping \( P^- \) constant at 0. We also vary \( T_{dur} \) within the range of \([0.5 \ 3]\) h with a 0.5-h increment, and adjust the number of clusters within the range of \([3 \ 36]\) with a 3-cluster increment. In Table V, the optimal parameter values indicate the ranges for generating multiple clusterings in Case 2. In particular, the values in bold represent parameters that denote maximum accuracy, which occurs \( P^+ > 0.8 \) kW and \( T_{dur} \geq 1.0 \) h. If \( P^+ \) and \( T_{dur} \) are too small, there will be very few data segments satisfying the low-power and minimum data length conditions. This renders the correlation calculation unreliable. Therefore, parameter ranges \( P^+ \) between \([0 \ 0.4]\) kW and \( T_{dur} \) between \([0 \ 0.5]\) h are removed from the subsequent simulations. The average computation times for parameter tuning on short, medium, and long feeders are 2.9 h, 3.6 h, and 39.5 h, respectively.

The findings indicate that the optimal number of clusters varies depending on the number of meters present on a circuit. As shown in Fig. 9, by maintaining \( P^- = 0 \) and increasing \( P^+ \) from 0.5 to 2, using 6 clusters yields the highest identification accuracy, with no further improvement achieved by increasing the number of clusters to 9. Analysis reveals that short feeders with tens of customers require only 6 clusters, medium feeders with a few hundred customers require 12 clusters, and long feeders with over 400 customers require 36 clusters. The findings suggest that circuits with longer laterals can benefit from grouping customers into more clusters. This is due to the potential weakening of voltage correlation between two meters on the same phase but supplied by different lateral circuits, as discussed in the type I circuit analysis in Section II-B. Consequently, dividing meters into additional groups enhances clustering accuracy.

We validate our algorithm performance through a two-round field verification process. In the first round, any mislabeled meters were corrected. Subsequently, we reran the algorithm to identify any remaining mislabeled meters. If such meters were found, utility engineers conducted a second field check to determine the accuracy of the identified mislabeled meters. This process will provide more thorough evaluation of the
TABLE V
PARAMETER SELECTION FOR REAL FEEDER DATA SETS

| Data type | Feeder no.  | Optimal parameter values |
|-----------|-------------|--------------------------|
|           | (No. of meters) | $P^+\sim kW = [0.2]$, $\Delta P^+=0.1$ | $T_{dur}[h] = [0.5,3]$, $\Delta T_{dur}=0.5$ | $3n^*=3[3,36]$, $an=1$ |
| Short     | 1 (33), 3 (73), 11 (24) | [0.8 / 1.2] | 1.0 / 1.5 | 6 |
| Medium    | 8 (173), 10 (131), 12 (108), 13 (137) | [1.3 / 1.7] | 2.5 / 3.0 | 18 |
| Long      | 2 (450), 4 (399), 5 (605), 6 (803), 7 (324), 9 (556) | [1.3 / 1.7] | 1.0 / 1.5 | 36 |

TABLE VI
CASE 1: PHASE IDENTIFICATION RESULTS AND PERFORMANCE COMPARISON

| Feeder No. | Phases in the utility records | Predicted phases | Round 1 | Round 2 | Accurancy computed in the first round. In practice, it is essential to establish a threshold that allows for the occurrence of a small number of “false negatives” rather than striving to eliminate all “false positives.” This is particularly important when utility engineers prioritize the identification of mislabeled meters for ensuring reliable operation.

As shown in Table VI, the proposed approach consistently outperformed the Spectral Clustering (SC) method [13] across all data sets. The window size for running SC was set to 4 days, and the number of clusters for each data set was determined as specified in Table V. This outcome demonstrates that using selected data segments, our algorithm captures voltage correlation between meters sharing the same phase better. In round 1, our method accurately predicts 93.7% of the original utility phase labels, with only 1.4% of “mislabeled” data found to be “false positive” following site verification.

The “false positives” meters are approximately 1% of the total meters, the majority of which are situated at the feeder’s far end and connected to the service transformer via lengthy secondary lines. These secondary lines exhibit higher impedance (as discussed in Section II-B), resulting in weaker voltage correlations between these meters and others supplied by the same transformer. This weakened correlation is the main reason for misclassification. Additional, a few “false positive” meters are characterized by exceptionally high power consumption, in which cases, only a few voltage segments are used for PCC calculation, leading to large variances and causing misclassification.

B. Case 2: Phase Identification Without Phase Labels

To generate multiple clusterings, we employ combinations of optimal parameter values from Table V. In this paper, we created 10 different clusterings by selecting 5 values for $P^+$, 2 for $T_{dur}$, and 1 for $3n^*$. The performance of the proposed method is compared with the Co-association Matrix Ensemble Clustering (CAM-EC) introduced in [14]. In CAM-EC, the window size is fixed at 4 days, and the number of clusters is determined as

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
follows: 3 and 6 clusters for short feeders, 6, 12, and 18 clusters for medium feeders, and 6, 12, 15, and 30 clusters for long feeders.

As shown in Table VII, the PBDS method outperformed CAM-EC in all cases. The identification accuracy of the PBDS method reached 100% for nearly half of the cases, whereas CAM-EC typically achieved accuracy below 100%. This difference in performance can be attributed to the fact that CAM-EC generates a cluster ensemble by combining different numbers of clusters, which can lead to inaccurate results unless the optimal number of clusters is determined for each specific data set. We also observed from the results that typically, feeders with a larger number of meters tend to have higher identification accuracy.

In the second round, although we can still identify some mislabeled meters, there is a sharp increase in the “false positive” rate. This observation suggests that in certain locations along a feeder, relying solely on APCC values may not be sufficient for achieving accurate phase identification, as discussed in the previous section.

C. Results of Pairing Identification

Table VIII reports the results of transformer-meter pairing identification. In stage 1, the numbers of flagged abnormal transformer-meter pairs are represented by ID1st and IDT2nd.
which are calculated using the average PCC values from (9) and (10), and the top-2 average PCC values from (11) and (12), respectively. \( ID^{2nd} \) and \( IDT^{2nd} \) indicate the number of meters excluded when checking \( APCC_i > e_{th} \). In stage 2, \( IDT^{2nd} \) and \( IDT^{2nd} \) represent the number of abnormal transformer-meter pairs after removing false positives using the PBDS method. Subsequently, utility engineers performed a GIS check and field verification on the flagged meters from both stages to identify potential false positives and false negatives. For Feeder 6, which has the highest number of DTs and meters, the computational time required was 75 minutes. This demonstrates the good scalability of the proposed algorithm, as it can handle larger systems effectively without any issues.

Overall, the 2-stage pairing identification algorithm achieves 52.6\% accuracy and a false positive rate of 47.4\%. In stage 1, using the PCC average (\( IDT^{1st} \)) and top-2 PCC average (\( IDT^{2nd} \)) methods without PBDS, approximately 3.4\% (136 out of 3,961 meters) and 1.6\% (64 out of 3,961 meters) are labeled as “abnormal” transformer-meter pairs, respectively. Applying the seasonal PCC check reduces the numbers to 117 (14\% reduction) and 58 (9.4\% reduction), respectively. PBDS’s impact on identification accuracy is evident in stage 2, as the subsequent method reduces false positives from 117 to 14 (88.0\% reduction) and from 58 to 17 (70.7\% reduction), respectively. This minimizes the need for field checks, alleviating utility concerns about high false positive rates.

Feeder 9 experiences multiple reconfiguration, resulting in the highest number of false positives and the lowest detection effectiveness (\( NV_2/NV_2 \)) among all the other feeders (refer to Table VIII). Additionally, the five remaining false positives are meters located at the far ends of distribution transformers (DTs), leading to weak voltage correlation with other meters on the same transformer.

V. CONCLUSION

In this paper, we applied power-band-based data segmentation (PBDS) in two meter topology identification algorithms: customer phase identification and transformer-meter pairing identification. In customer phase identification, we utilized the low power band condition to select voltage segments. In transformer-meter pairing identification, we employed a two-stage method where in the second stage, PBDS is used for reducing false positives by excluding meters with extremely low power consumption and obtaining voltage segments with larger correlation gaps. This step played a crucial role in distinguishing between meters supplied by the same transformer and those supplied by different transformers on the same phase. Through simulation on thirteen real feeders, our proposed algorithm outperformed existing approaches, significantly improving the accuracy of meter topology identification without introducing significant computational complexity.

REFERENCES

[1] J. S. Guerrero-Prado, W. Alfonso-Morales, and E. F. Caicedo-Bravo, “A data analytics/Big Data framework for advanced metering infrastructure data,” Sensors, vol. 21, no. 16, 2021, Art. no. 5650.

[2] D. Burns et al., “2020 assessment of demand response and advanced metering,” Federal Energy Regulatory Commission, Tech. Rep., Dec. 2020. [Online]. Available: https://www.ferc.gov/media/2021-assessment-demand-response-and-advanced-metering.

[3] T. Shen, M. Jakisić, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, “Three-phase AC system impedance measurement unit (IMU) using chirp signal injection,” in Proc. 28th Annu. IEEE Appl. Power Electron. Conf. Expo., 2013, pp. 2666–2673.

[4] M. H. Wen, R. Arghandeh, A. V. Meier, K. Poolla, and V. O. Li, “Phase identification in distribution networks with micro-synchrophasors,” in Proc. IEEE Power Energy Soc. Gen. Meeting, 2015, pp. 1–5.

[5] V. Arya et al., “Phase identification in smart grids,” in Proc. IEEE Int. Conf. Smart Grid Commun., 2011, pp. 25–30.

[6] V. Arya, T. Jayram, S. Pal, and S. Kalyanaraman, “Inferring connectivity model from meter measurements in distribution networks,” in Proc. 4th Int. Conf. Future Energy Syst., 2013, pp. 173–182.

[7] H. Pezeshki and P. J. Wolfs, “Consumer phase identification in a three phase unbalanced LV distribution network,” in Proc. 3rd IEEE PES ISGT Europe, 2012, pp. 1–7.

[8] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, “Phase identification of smart meters by clustering voltage measurements,” in Proc. Power Syst. Compt. Conf., 2018, pp. 1–8.

[9] R. Mitra et al., “Voltage correlations in smart meter data,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2015, pp. 1999–2008.

[10] W. Wang, N. Yu, B. Fogg, J. Davis, and J. Li, “Phase identification in electric power distribution systems by clustering of smart meter data,” in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl., 2016, pp. 259–265.

[11] Z. S. Hosseini, A. Khodaei, and A. Paaso, “Machine learning-enabled distribution network phase identification,” IEEE Trans. Power Syst., vol. 36, no. 2, pp. 842–850, Mar. 2021.

[12] B. Fogg and N. Yu, “Improving supervised phase identification through the theory of information losses,” IEEE Trans. Smart Grid, vol. 11, no. 3, pp. 2337–2346, May 2020.

[13] L. Blakely, M. J. Reno, and W.-C. Feng, “Spectral clustering for customer phase identification using AMI voltage timeseries,” in Proc. IEEE Power Energy Conf. Illinois, 2019, pp. 1–7.

[14] L. Blakely and M. J. Reno, “Phase identification using co-association matrix ensemble clustering,” IET Smart Grid, vol. 3, no. 4, pp. 490–499, 2020.

[15] H. P. Lee et al., “A novel data segmentation method for data-driven phase identification,” in Proc. IEEE Power Energy Soc. Gen. Meeting, 2022, pp. 1–5.

[16] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, “Identifying topology of low voltage distribution networks based on smart meter data,” IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 5113–5122, Sep. 2018.

[17] W. Luan, J. Peng, M. Maris, J. Lo, and B. Harapnuik, “Smart meter data analytics for distribution network connectivity verification,” IEEE Trans. Smart Grid, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.

[18] J. D. Watson, J. Welch, and N. R. Watson, “Use of smart-meter data to determine distribution system topology,” J. Eng., vol. 2016, no. 5, pp. 94–101, 2016.

[19] Y. Weng, Y. Liao, and R. Rajagopal, “Distributed energy resources topology identification via graphical modeling,” IEEE Trans. Power Syst., phase II, no. 4, pp. 2682–2694, Jul. 2017.

[20] L. Blakely and M. J. Reno, “Identifying errors in service transformer connections,” in Proc. IEEE Power Energy Soc. Gen. Meeting, 2020, pp. 1–5.

[21] T. A. Short, “Advanced metering for phase identification, transformer identification, and secondary modeling,” IEEE Trans. Smart Grid, vol. 4, no. 2, pp. 651–658, Jun. 2013.

[22] C. Ye, C. Jiming, J. Hao, G. Yajuan, W. Jiang, and T. Haibo, “Two-stage topology identification method for distribution network via clustering correction,” in Proc. IEEE Innov. Smart Grid Technol.-Asia, 2019, pp. 280–284.

[23] L. Blakely and M. J. Reno, “Identification and correction of errors in pairing AMI meters and transformers,” in Proc. IEEE Power Energy Conf. Illinois, 2021, pp. 1–8.

[24] T. J. Blalock, “The first polyphase system: A look back at two-phase power for AC distribution,” IEEE Power Energy Mag., vol. 2, no. 2, pp. 63–66, Mar./Apr. 2004.

[25] J. Yadav and M. Sharma, “A review of k-mean algorithm,” Int. J. Eng. Trends Technol., vol. 4, no. 7, pp. 2972–2976, 2013.

[26] H. Lee, “Machine learning applications for smart meter data analysis,” 2023. [Online]. Available: https://repository.lib.ncsu.edu/items/cd8755e9-8334-4f24-9d39-7d75580928a
Han Pyo Lee (Senior Member, IEEE) received the B.S. degree in electrical & electronic engineering from Yonsei University, Seoul, South Korea, in 2008, the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2020, and the Ph.D. degree in electrical engineering from the Future Renewable Electric Energy Delivery and Management (FREEDM) Systems Center, North Carolina State University, Raleigh, NC, USA, in 2023. His research interests include machine learning of distribution system planning models, data-driven modeling and control of energy systems, and voltage control for renewable energy-integrated power systems.

PJ Rehm received the Bachelor of Science degree in mechanical engineering from North Carolina State University, Raleigh, NC, USA. He is currently the Director of Lifecycle Operations at FlexGen, a leading battery storage integration and software company. He leads the team responsible for maximizing uptime and availability for battery energy storage assets. Previous to this role, he was the Vice President, Grid Innovation and Safety at ElectriCities of NC, where he spent more than 12 years focused on helping member utilities advance and deploy utility technologies. He helped members with AMI system deployments, demand response programs, renewable programs, customer engagement solutions, and more.

Matthew Makdad received the B.S. degree in electrical engineering from The University of Pennsylvania, Philadelphia, PA, USA. P.E. has been the Engineering Supervisor at Industry Partner New River Light and Power in Boone since 2017. He oversees planning and operations which include the line workers, metering shop, and mapping functions. He received his NC Professional Engineering license in 2003. He has been integral in reviewing the project analysis and performing data field validation.

Edmond Miller (Associate Member) biography and photograph not available at the time of publication.

Ning Lu (Fellow, IEEE) received the B.S. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1993, and the M.S. and Ph.D. degrees in electric power engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1999 and 2002, respectively. She is currently a Professor with the Electrical and Computer Engineering Department, North Carolina State University, Raleigh, NC, USA. Her research interests include modeling and control of distributed energy resources, microgrid energy management, and applying machine learning methods in energy forecasting, modeling, and control.