This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

Reduced Voltage-Dependency by Categorical Location Information and Distance Along Street Metric for Meter-Transformer Mapping in Distribution Systems

Bilal Saleem, *Student Member, IEEE,* Yang Weng, *Senior Member, IEEE,* Vijay Vittal, *Life Fellow, IEEE*

***Abstract*—Deep penetration of distributed energy resources (DERs) and electric vehicles (EVs) introduce benefits but may cause the overloading of service transformers in distribution networks. Such situations require real-time transformer loading information, where an accurate mapping between smart meters and distribution transformers is a prerequisite, e.g., summing the downstream smart meter consumptions. Due to arbitrary curvature in streets, we propose to employ a density-based clustering based on voltage magnitudes (continuous) and street name (categorical) information. However, a density-based approach may only be able to localize a meter to a street segment. Hence, we use a second-stage spectral clustering with distance along the street (DAS), a novel feature, to obtain meter clusters, each with a common parent transformer. For mapping transformers to meter clusters, we use the nearest cluster center approach based on the location since voltage measurements may not be available at transformers. Moreover, we provide a theoretical guarantee for such an approach. Finally, we illustrate the usefulness of the proposed algorithm on long streets, which is a challenging scenario due to many possible incorrect combinations of meter-transformer mapping. The proposed algorithm has been tested on modified IEEE 8-, 69-, 123-bus test systems and real distribution feeders from a utility in the Southwestern United States, demonstrating outstanding performance.**

I. Introduction

Distributed energy resources (DER) and electric vehicles (EV) are environmentally beneficial. However, as their penetration increases, they may cause service transformer overload since their incorporation increases the forward and reverse power flow via those transformers. Long-term overloading of transformers causes insulation deterioration and shorter lifespans. Additionally, because it is expensive to meter every service transformer, many utilities do not have the real-time loading information of transformers. For example, the cost of remote sensing on a distribution transformer is around $1,000$ U.S. dollars [1]. Moreover, the cost of a Fluke 435-II meter [2] is around $9,000$ U.S. dollars. Furthermore, many utilities do not have detailed existing distribution system topologies [3]. Thus, it is hard to add the downstream meter consumptions to determine how much their parent transformers are loaded [4].

One idea is to replicate the transmission grid's methods for determining topology. However, transmission grid techniques [5]–[9] rely on established telemetry and rare topology changes, making them unsuitable for distribution grids. However, utilities can exploit the huge data from the implementation of advanced metering infrastructure (AMI) [10]–[13]. In order to reconstruct the system topology using AMI and micro-synchrophasor data, research has been done to develop algorithms that use voltage magnitude correlations [14]. For instance, [15] estimates system topology using voltage magnitudes but assumes that all lines have equal inductance-to-resistance ratios per unit length. In addition, the Chow-Liu method is used in [16] to determine radial topology. References [17], [18] predict the topology and line characteristics using historical data of real and reactive power, voltage magnitudes, and voltage angle. Voltage magnitudes at each system node are necessary for these procedures. However, voltage magnitude sensing may not be available at poles and transformers due to the enormous number of buses in distribution networks.

Some efforts are dedicated to the distribution grid [19]–[21]. However, they require the locations of all switches or the most likely topology, which may be unavailable due to the vast spread of distribution lines [22]. Other efforts even require impedance [23], which may be unavailable in the secondary distribution grids. References [24], [25] primarily use AMI voltage data to recover the mapping between smart meters and service transformers. Such a mapping can help utilities estimate the loading of distribution transformers, e.g., the utility can sum the powers flowing through the daughter smart meters to estimate the power flowing through their parent transformer.

However, analyses based solely on voltage magnitudes may not be accurate because smart meters that are geographically dispersed and unconnected can have similar voltage magnitude profiles due to similar neighborhood consumption profiles. Furthermore, high photovoltaic (PV) penetration makes the meter consumptions and, therefore, the voltage magnitude similar, even if the meters belong to different transformers. Such similarity may cause problems for algorithms that consider voltage magnitude only. However, considering information in addition to the voltage magnitudes, e.g., location information, can prevent wrong results by considering only voltage data. Another challenge is that the meters close to different transformers have similar voltage magnitudes in the case of similar net consumptions flowing through transformers, and hence, they tend to be clustered together. This occurs because the impedance of primary conductors is negligible when referred to the transformer's secondary side.

Researchers also focus on using geographical information to aid voltage magnitude-based methods, but geographical information has not been effectively used. For example, many voltage magnitude-based methods implicitly utilize geographic information system (GIS) data only for selecting meters belonging to a geographical area and then use their voltage

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

magnitude-based method on the selected meters. Moreover, [26], [27] use the GIS information of meters only to estimate voltage magnitudes at the point of connection by using average energy and estimating the distance from a house to the nearest pole. Reference [25] uses both voltage and geographical information. However, it puts equal emphasis on both pieces of information, which can not be altered. Furthermore, using geographical information directly without preprocessing can create problems. For instance, both sides of a street are typically supplied by the same transformer. However, without preprocessing the GIS information, the results may be incorrect. Based on the above discussion, one can raise the following questions. How should one use the GIS information to aid voltage magnitude-based clustering effectively? Moreover, as measurements are unavailable at the transformers, can there be a guarantee for matching transformers to meter clusters? Finally, meters close to different transformers have similar voltage magnitudes. How does one distinguish between them? To resolve these issues, this paper has three contributions listed below:

1) We propose to use the GIS information in the most effective manner. For example, we first divide the data into streets, using the street categorical information along with the voltage magnitudes via density-based clustering [28]. Next, we use the geocoded street center as a categorical feature, which has the same value for all meters in a particular street. Using voltage magnitudes in conjunction with the street categorical information also takes care of meters around the junction of two streets. We designed density-based clustering as the first stage due to arbitrarily shaped streets, and flexibility in the amount of input data (parameters constant). Also, density-based clusters contain whole numbers of transformer secondary circuits, e.g., a transformer secondary is not split between the density-based clusters.
2) Next, we compute the distance along the street (DAS) as a feature for the meters in the street, considering the first street meter as the reference point for computing the distance. We use DAS along with voltage magnitudes as features for spectral clustering [29] to cluster meters supplied by the same transformer together. We designed the algorithm utilizing DAS as it is more useful than the usual Euclidean or Haversine metrics because distribution lines are usually laid along the streets. Moreover, the DAS metric treats both sides of a street equally and is easy to compute. Due to its superior performance, we choose spectral clustering [25]. Spectral clustering needs the number of clusters (transformers), so we devised a mechanism to identify the number of transformers in a density-based cluster via discrete optimization.
3) Once the meter clusters are obtained, the next challenge is identifying the parent transformers of the clusters. As we assume no measurements on transformers, we only use the location information to identify the parent transformers. Furthermore, we propose a proof of a novel probabilistic guarantee for identifying the parent transformer using location information only. Such a guarantee is possible by utilizing the categorical information and the DAS metric. Another challenge is that long streets have many transformers, i.e., many incorrect possible choices, in contrast to only one correct choice. Moreover, similar neighborhood consumptions are more likely to occur in the case of more transformers, resulting in similar voltage magnitudes for different transformer secondary meters. Hence, we distinguish such meters using our location information strategy described above. Our method only needs smart meter data and no additional metering device.

Using such an approach, we can show benchmark and real system scenarios not solvable by the earlier methods. Furthermore, we have used diverse test-system topologies to compare our proposed method. For example, we validate using the IEEE 8-bus system, IEEE 69-bus system, IEEE 123-bus system, and our partner utility's high PV penetration distribution feeder with around $1,800$ customers. The results demonstrate that the proposed method accurately segments the smart meter data to identify meter-transformer mapping.

Our method is not just useful for identifying transformer loading, but it also has an influence on classic estimation problems in power system research. For example, we show how our method can influence state estimation, topology estimation, and line parameter estimation problems. Next, we also discuss how voltage measurement errors can affect our method and the estimation problems.

The rest of the paper is organized as follows: Section II presents the problem formulation. Section III introduces the location data preparation. Section IV shows how to merge heterogeneous data to obtain meter clusters supplied by the same transformer. Section V provides proof of the guarantee. Section VI provides meter-transformer mapping identification for long streets. Section VII discusses the influence on other estimation problems. Section VIII discusses the operating conditions. Section IX validates the idea numerically, and Section X concludes the paper.

II. Problem Formulation

To formulate the proposed algorithm, we assume that the time series voltage magnitude information is available for N smart meters $\mathbf{x}^1, \cdots, \mathbf{x}^N$. For instance, the addresses for N smart meters $\mathbf{a}^1, \cdots, \mathbf{a}^N$ are stored as rows in dataset A. The latitude-longitude pairs in radians for N smart meters $\mathbf{l}^1, \cdots, \mathbf{l}^N \in \mathbb{R}^{2\times 1}$ are stored as row vectors in matrix $L \in \mathbb{R}^{N\times 2}$, where $\mathbb{R}$ represents the set of real numbers. The street centers' global positioning system (GPS) coordinates $\mathbf{h}^1, \cdots, \mathbf{h}^N \in \mathbb{R}^{2\times 1}$ are obtained by geocoding the street information and stored as row vectors in the street location matrix $H \in \mathbb{R}^{N\times 2}$. The voltage magnitude time-series with T timeslots for N smart meters $\mathbf{v}^1, \cdots, \mathbf{v}^N \in \mathbb{R}^{T\times 1}$ are stored as row vectors in matrix $V \in \mathbb{R}^{N\times T}$. In addition to smart meters, we assume that there are k transformers forming k clusters of smart meters in the distribution grid. Also, transformer locations are available. C_j represents a set of indices of all smart meters in the j^{th} cluster. A smart meter $i \in \{1, \cdots, N\}$ is uniquely present in a cluster $j \in \{1, \cdots, k\}$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

that is supplied by a common transformer. There exists a many-to-one mapping $f : i \to j$.

For correlating these variables, a distribution system is characterized by buses $\mathcal{V} = 1, 2, \cdots, N$ and by branches $\mathcal{E} = (i, i'), i, i' \in \mathcal{V}$. The voltage measurement data at bus i and time t is represented as the magnitude of the instantaneous voltage at bus i in per-unit $|v_i(t)| \in \mathbb{R}$. The voltage magnitude readings taken by the meters in $\mathbf{v}^i$ are root mean square values over a time period based on the rate of utility collection. In this study, it is assumed that there are various signal processing techniques to denoise the data prior to our analysis. The scenario being considered is:

- Problem: identify smart meter to transformer connectivity
- Given: smart meter voltage magnitude data V, the smart meter address dataset A, and the smart meter street GPS coordinates L.
- Find the M.T. mapping $f : i \to j$.

III. GIS Data Preparation

Previous methods for identifying meter-transformer mapping in the literature primarily use voltage magnitudes. However, voltage magnitudes are useful, but they do not give complete information on meter-transformer mapping. For example, it so happens that distant meters have similar voltage magnitudes due to similar neighborhood consumption profiles even when they don't share a common distribution transformer. In fact, meters in different cities can have similar voltage profiles, which is a source of error for methods that use voltage information only. Therefore, the identification of meter-transformer mapping needs information from multiple sources to increase accuracy. For instance, location data can be classified into three types: nodes, edges, and polygons. A node is characterized by a latitude-longitude pair. An edge is defined by two nodes as endpoints, whereas a polygon is defined by multiple nodes as the corners. Below we classify the GIS information in power systems into three classes to describe how we use such information for meter-transformer mapping.

A. Nodal Information

Nodal information comprises the GPS coordinates of the distribution system devices, e.g., transformers, poles or underground manholes, and meters. The following metric is helpful for the identification of meter-transformer information.

Direct distance between two nodes (Haversine): The shortest distance between two points on a spherical surface is the Haversine distance. The Haversine distance is given by a simple formula, assuming the planet Earth is a sphere. For example, let $\mathbf{l}^1, \mathbf{l}^2 \in \mathbf{L}$ be two latitude-longitude pairs in radians. The distance between them on planet Earth's surface is given using the Haversine formula. Let $A_1 := \sin^2\left(\frac{\Delta l_1}{2}\right)$ and $A_2 := \sin^2\left(\frac{\Delta l_2}{2}\right)$. Then,

$$d_{Hav}(\mathbf{l}^1, \mathbf{l}^2) = 2R_E \cdot \arcsin\left(\sqrt{A_1 + \cos(l_1^i)\cos(l_1^j)A_2}\right), \tag{1}$$

where $R_E = 6371$ km is the radius of planet Earth.

B. Edge-related Information

Although nodal information is important for GIS-based identification, additional information is needed. Specifically, GPS coordinates do not provide edge-related information, e.g., streets, street blocks, or street shape. Edge-related information is helpful since distribution system lines are usually laid along streets. As more information is better, we consider both continuous and categorical information.

1) Categorical Street Information: Usually, the meters are not supplied by transformers located on different streets. For instance, the meters and their parent transformers are typically located on the same street. This is especially true as making the low voltage distribution lines cross wide streets is costlier. For overhead distribution, taller and more expensive poles would be needed to cross wide streets due to the clearance limit. For underground distribution, it would be more expensive to maintain when lines frequently cross streets.

Hence, a categorical street identifier is needed for recovering meter-transformer mapping. For the identifier, one idea is to use a unique random identifier for each street. It can be used to construct a test case comprising n-nearest streets for each meter in a central street. However, such a street identifier does not carry useful information about the street concerning the meter in consideration. A better idea is to use the latitude and longitude of the street center $\mathbf{h}$ as a unique identifier for each street. Such an identifier considers the relative position of the street with the meter in consideration. Apart from the categorical information, a street identifier is also needed to formulate the following edge-related distance metric.

2) Distance Computation along the Streets (DAS): One way to compute the distance between two nodes is to compute DAS. Overhead and underground power distribution lines and devices usually follow streets. DAS is a metric, which is thus more useful than the Haversine distance for meter-transformer mapping. For example, such a metric is related to the length of lines, which should be kept shorter to reduce losses. DAS can be obtained using the A* (A-star) search algorithm [30]. For instance, first, a graph is constructed by considering each address and street junction stored as single points. Two points represent a straight street. Irregular streets are stored as a series of points, such that straight line segments drawn between successive points appear as a curve. This series of points is called a polyline.

After the graph construction, it is searched using the A* algorithm [30]. The algorithm estimates the cost of reaching the goal from the source and explores the paths that minimize the total cost via best-first search [31]. Online map services also use such an algorithm. Online maps services store location information for all houses and streets in a GIS, in the form of points with latitude, longitude, and altitude. DAS can be easily computed using an online-maps service, e.g., Google Maps, via an application programming interface (API). We use $d_{DAS}(\mathbf{a}^1, \mathbf{a}^2)$ to refer to the DAS in km between two addresses $\mathbf{a}^1, \mathbf{a}^2$.

C. Polygon-related Information

Polygons can identify the boundaries of houses, zip codes, cities, and other such relevant information. However, such data

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

is not very useful for power distribution systems since such features are already included in the edge-related information. For example, online map services already process such information to identify the directions and DAS.

IV. Merging Heterogeneous Data for Meter-Transformer Mapping Identification

In the previous section, we identified three metrics based on geographical information, i.e., streets, direct distance, and distance along streets. This section shows how to utilize the information of the three distance metrics and information from voltage magnitudes. Generator buses in a transmission system have fixed voltages and are usually modeled as PV buses. However, in distribution system, the buses associated with distributed energy resources (DER) are usually modeled as PQ buses.

For example, there are three different types of control variables for DERs: voltage control, current control, and P-Q control. The associated bus for DER with voltage controller is modeled as a PV bus (fixed voltage). Moreover, the associated bus for DER units that have either current controller or P-Q controller is modeled as a PQ bus [32]. However, according to the IEEE P1547 Standard for Interconnection [33]. DER's attempt to control distribution system voltage may clash with utility voltage regulation plans already in place to change the voltage at the same or a nearby site [34]. Thus, it is not advised to use DERs with a voltage regulator. Therefore, in this study, we model the DER bus as PQ bus rather than a PV bus.

As discussed in Section I, density-based clustering and spectral clustering are utilized to characterize clusters. The clustering methods are detailed below.

A. Density-Based Clustering with Categorical Information

Our proposed algorithm incorporates the street-crossing constraint intelligently. For instance, the algorithm puts more weight on meter-transformer mapping in the same street. However, for cases with a high voltage magnitude similarity with neighboring street meters, such meters are connected to transformers of the neighboring street.

The proposed metric for density-based clustering contains both continuous and categorical street information. For example,

$$d(\mathbf{x}^1,\mathbf{x}^2) = \gamma d_{categ}(\mathbf{x}^1,\mathbf{x}^2) + d_{cont}(\mathbf{x}^1,\mathbf{x}^2), \quad (2)$$

$$= \gamma d_{Hav}(\mathbf{l}^1,\mathbf{l}^2) + d_{volt}(\mathbf{v}^1,\mathbf{v}^2), \quad (3)$$

where $categ$ stands for categorical information and $cont$ stands for continuous information. d_{cont} represents the voltage magnitude-distance of the two points $\mathbf{x}^1, \mathbf{x}^2$. We consider the voltage magnitude-distance formed by mutual information as $\frac{1}{I(\mathbf{v}^1,\mathbf{v}^2)}$. For the categorical information, we consider the distance between two points on the same street to be zero. Moreover, the distance between two points on different streets is the distance between the street centers $\mathbf{l}^1, \mathbf{l}^2$. γ is the coefficient for merging categorical and continuous distance.

From Eq. 3, since $d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$ and d_{volt} are added together, they should be comparable. For instance, $d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$ is a physical distance on the surface of Earth, and it is measured in kilometers. However, $d_{volt}(\mathbf{v}^1,\mathbf{v}^2)$ is defined as $\frac{1}{I(\mathbf{v}^1,\mathbf{v}^2)}$, which is unit-less. So, γ has a unit per km. Furthermore, the value of γ can be estimated as the mean of the quantity $d_{volt}(\mathbf{v}^1,\mathbf{v}^2)/d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$ for a sample of meters. The allowed range of γ can be written as $d_{volt}(\mathbf{v}^1,\mathbf{v}^2)/d_{Hav}(\mathbf{l}^1,\mathbf{l}^2) \pm \zeta$, where ζ can be defined as the standard deviation of the quantity $d_{volt}(\mathbf{v}^1,\mathbf{v}^2)/d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$. Since $d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$ is measured in km, usually $d_{volt}(\mathbf{v}^1,\mathbf{v}^2)/d_{Hav}(\mathbf{l}^1,\mathbf{l}^2)$ is greater than 1.

1) Notion of Density in Euclidean Space: Consider a two-dimensional XY space to define the concept of density in a high-dimensional space. Consider any two points without the loss of generality. $(l_1^1, l_2^1), (l_1^2, l_2^2) \in \mathbb{R}^2$ in a 2-D space. In two dimensions, the Euclidean distance is $[(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{\frac{1}{2}}$. If we set the distances to be less than $epsilon$, we get: $[(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{\frac{1}{2}} < \epsilon$. Squaring both sides of the last equation yields: $(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2 < \epsilon^2$. The equation represents a circular disk with radius ϵ and center at the point (l_1^2, l_2^2). For high dimensions, if the count of data points in the sphere of radius $epsilon$ around a data point exceeds $minPoints$, the algorithm considers the central data point as a core point. Due to the planet Earth's spherical shape, utilizing Euclidean distance is incorrect, so we employ the Haversine distance, which gives the distance on the Earth's surface considering the spherical path.

2) Notion of Density in the Combined Space: To define the concept of density in the combined continuous and discrete data with distance metric shown by (2), we need to define the neighborhood of a datapoint.

Definition 1: (ϵ-neighborhood of a point) The ϵ-neighborhood of a datapoint $\mathbf{x}^1$ denoted by $N_r(\mathbf{x}^1)$, is defined by

$$N_r(\mathbf{x}^1) = \left\{\mathbf{x}^2 : d_{Hav}(\mathbf{x}^1,\mathbf{x}^2) < \epsilon\right\}, \quad (4)$$

where $d_{Hav}(\mathbf{x}^1,\mathbf{x}^2)$ is defined by (2).

The ϵ-neighborhood of a point is a notion of the density of points. If $N_r(\mathbf{x}^1) > minPoints$ then $\mathbf{x}^1$ is a *core point*. Datapoints at a cluster's boundary may not qualify as core points. For such points, we cluster them with a core point if they are in the ϵ-neighborhood of a core point.

The density-based clustering stage segments the big utility dataset into substreet clusters, which are processed by the stage described below to recover the meter-transformer mapping.

B. Identify Meter-Transformer Mapping using DAS with Voltage Magnitude Information via Spectral Clustering

Spectral Clustering has been utilized in [25] to identify meter-transformer mapping under two challenging scenarios, large distances between a meter and its parent transformer or high similarity of a meter's consumption pattern to a non-parent transformer's meter. Such a method is based on similar voltages for meters supplied by a distribution transformer. The voltages within a distribution transformer secondary are similar due to the voltage drop across the transformer. However, such an application does not utilize the DAS, which is needed in the case of long streets. For example, such a method can not identify the correct meter clusters in the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

case of long streets where it is difficult to choose the correct transformer for a meter given the large number of transformers to choose from. Therefore, location information is important. Moreover, using location information directly is not logically correct since it discriminates between the two sides of a street. However, based on the distribution system planning, meters on both sides of a street are supplied by the same transformer, except in exceptional scenarios, e.g., a high power-consuming commercial consumer in a residential area that may be routed to a nearby transformer.

Given a sub-street cluster from the density-based clustering step. Let the sub-street cluster have N smart meters, the voltage magnitude time series $\mathbf{v}^1, \cdots, \mathbf{v}^N$ and the distance along the street information $0 = d^1, d^2, \cdots, d^N$, where $d^1 = 0$ is the reference point. Spectral clustering clusters them into k transformer secondary clusters as follows:

1) Consider the voltage magnitude affinity matrix $M_v \in \mathbb{R}^{N\times N}$ as the Pearson Correlation Coefficient Matrix [25].
2) Consider the distance affinity matrix $M_d \in \mathbb{R}^{N\times N}$ with element $m_{d,ij} = \left(1 + \beta d_{DAS}\left(\mathbf{a}^1, \mathbf{a}^2\right)\right)^{-1}$, where β is of the order of 100.
3) Define D_v to be a diagonal degree matrix with elements $d_{v,ii} = \sum_{j=1}^{N} m_{v,ij}$, where $m_{v,ij}$ are the elements of M_v. Similarly, D_d is the diagonal degree matrix with elements $d_{d,ii} = \sum_{j=1}^{N} m_{d,ij}$, where $m_{d,ij}$ are the elements of M_d.
4) Construct the combined affinity matrix $M = \alpha M_v + (1-\alpha)M_d$ and the combined diagonal degree matrix $D = \alpha D_v + (1-\alpha)D_d$. Moreover, construct the graph Laplacian matrix $\mathcal{L} = D - M$.
5) Select the number of groups k as the no. of transformers.
6) Find $\mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(k)}$, the right eigenvectors corresponding to k smallest eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$ of $\mathcal{L}$. Since $\mathcal{L}$ is real and symmetric, the left eigenvectors are simply the transpose of the right eigenvectors. Form the matrix $U = \left[\mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(k)}\right] \in \mathbb{R}^{N\times k}$. Since eigenvectors are orthogonal to each other, doing so will further distance the points belonging to different clusters.
7) Treat each row of U as a point in $\mathbb{R}^k$ and cluster via $k-$means++ [25]. Due to orthogonalizing using the eigenvectors, the data points belonging to separate clusters are almost orthogonal to each other, i.e., they have approximately right angles at the origin with respect to (w.r.t.) each other, so that $k-$means++ can cluster well.
8) For U's rows assigned to cluster C, the original corresponding points s_i are present in cluster C [35].

C. Number of Transformers Within Density-Based Clusters

The density-based (DB) clusters cannot directly identify the meter-transformer mapping. It is because such a method does not utilize the information on the number of transformers. However, as the density-based clustering uses the street categorical information and the voltage magnitude information, the clusters comprise whole numbers of transformer secondary distribution circuits. In other words, a transformer secondary circuit is not split between two or more density-based clusters.

Furthermore, by utilizing the categorical street information, the size of the DB clusters is not larger than the size of the streets. Hence, we can utilize the distance along the street for the next step. However, the number of transformers supplying a DB cluster is unknown, which is required. For instance, if we know the number of transformers supplying a DB cluster, we can further decompose the DB clusters to obtain subclusters with a single transformer so that the subclusters minimize the sum of inter-subcluster similarity matrix elements M_{ij}.

Density-based spatial clustering of applications with noise (DBSCAN [36]) is also efficient in handling outliers. Therefore, subsequent Spectral Clustering will improve the result further. In addition, the final k-means++ step of Spectral Clustering has a higher chance of convergence to the global minimum if we have fewer clusters. However, for more clusters, k-means++ is more likely to be trapped in a local minimum. Therefore, we choose DBSCAN to precede Spectral Clustering for reducing the number of clusters the final k-means++ step needs to process at a given time.

The combined algorithm requires the hyper-parameters of both Spectral Clustering and DBSCAN but performs better than both algorithms working individually. The algorithm specifically needs the total number of clusters m, the radius ϵ for computing density, and the minimum points $minPoints$ inside the radius for density calculation. Let m be the number of transformers feeding smart meters for the meter-transformer mapping. Fig. 1 depicts the combined algorithm's flowchart.

The DBSCAN method with the parameters ϵ and $minPoints$ is used to cluster the data in the first stage. p is the number of clusters DBSCAN provides depending on the two parameters. In the second stage, the p DBSCAN clusters are further divided into m clusters by applying Spectral Clustering to each DBSCAN cluster. In order to apply Spectral Clustering, it is important to determine the number of clusters. Determination of the number of clusters requires considering the number of ways to choose p groups from m transformers.

We begin with the scenario of each segment having at least one transformer. For example, if we draw m transformers in a line, we get $m-1$ spaces between the transformers. Hence, we choose $p-1$ spaces to form p segments. So, the total number of choices is $\binom{m-1}{p-1}$. We can represent the number of choices as the number of solutions to an equation. For instance, if we consider E_i as the number of transformers in segment i, then $E_1 + \cdots + E_p = m$, where $E_i > 0$. So, we have $\binom{m-1}{p-1}$ solutions. We will pick the solution where the m subclusters minimize the the sum of inter-subcluster similarity matrix elements M_{ij}. Fig. 1 depicts the combined algorithm's flowchart.

Similarly, if we also consider the possibility of having segments without transformers, then the number of ways we can form p segments from m transformers can be represented as the number of solutions of the equation $E_1 + \cdots + E_p = m$, where $E_i \geq 0$. Therefore, if we define $F_i = E_i + 1$, then we get $F_1 + \cdots + F_p = m + p$, where $F_i > 0$. This becomes the same scenario as before. So, the number of solutions to this equation is $\binom{m+p-1}{p-1}$, which is the number of ways we can form p segments from m transformers by considering segments may have zero transformers. However, a cluster is

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

useless without a transformer. As a result, the minimal value of k is 1, as considered before.

Consider the scenario when we know there are $m = 5$ transformers overall in our sample data and DBSCAN offers $p = 3$ clusters. Thus, there are six ways to divide five clusters into three groups since $\binom{5-1}{3-1} = \binom{4}{2} = 6$ k-sets. The ways are as follows: (1,2,2), (2,1,2), (2,2,1), (1,1,3), (1,3,1), and (3,1,1). As a result, we use Spectral Clustering to partition each DBSCAN cluster into two and three subclusters. The option with the smallest sum of inter-subcluster similarity matrix elements, M_{ij} is finally chosen.

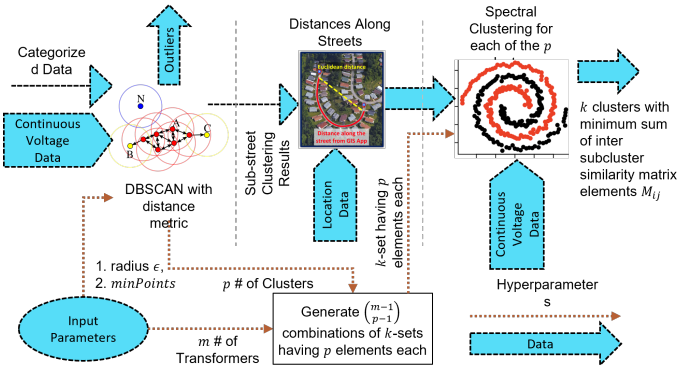


Fig. 1. Block diagram of the new combined algorithm.

D. Identification of Loops for Meter-Transformer Mapping

The earlier method works fine for tree distribution, meshed primary distribution, and mesh structures within one transformer secondary. For completely meshed secondary distribution, all transformers' secondary circuits are connected together. In such a case, the meter-transformer mapping is trivial. However, the method described previously could not work in a looped structure, e.g., when the secondary circuits of two transformers are interconnected. To make it work, we need to identify the clusters that have transformer secondaries connected. For example, knowing such information indicates one cluster instead of two.

We devise a metric based on the sorted similarity matrix by spectral clustering to determine if there is a loop in the secondary distribution network. For example, in Fig. 2, to check if the secondaries of transformers corresponding to clusters C_1 and C_2 are connected, we devise a metric $confidence = C_1 + C_2 - 2I_1$, where C_i is the mean of the diagonal submatrix and I_i is the mean of the off-diagonal submatrix. In the case of separate secondaries, the similarity of inter-cluster I will be lower, and therefore, the $confidence$ metric will be greater than 0.5. Similarly, in the case of connected secondaries, the similarity of inter-cluster I will be higher, and therefore the $confidence$ metric will be lower than 0.5. Therefore, we can use 0.5 as a cut-off value to determine if both transformers' secondaries are connected.

E. Apply Clustering Results to Identify Smart Meter - Distribution Transformer Mapping

In Section IV-C, we clustered smart meters so that each cluster is supplied by one distribution transformer. The next

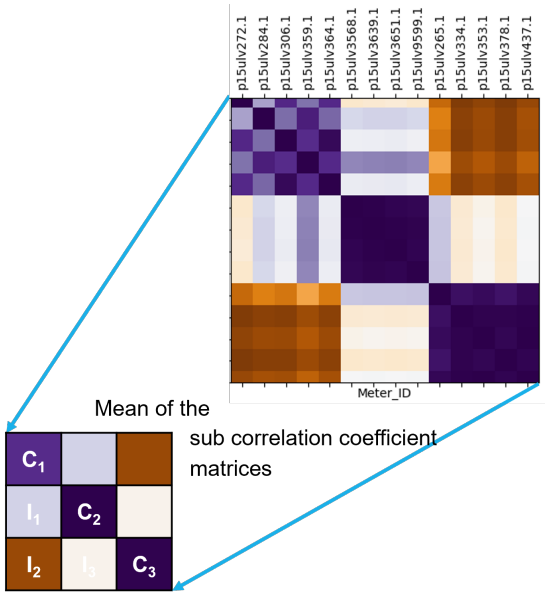


Fig. 2. Devise a metric to determine looped subsystems.

step is to identify the parent transformers for the smart meter clusters. As there is no monitoring available on distribution transformers, the only available information is the location of the transformer. Therefore, we match the meter clusters to their parent transformer using location information. The steps for such a match problem are given below.

1) Using meter locations, identify the mean (center) location for each meter cluster.
2) Find the nearest transformer to each meter cluster center and assign the cluster to the transformer.

It is better to assign the nearest transformer to a meter cluster center than vice versa, as meter clusters usually do not overlap. For example, it is easier for utilities to route electrical wires and cables, so that transformer secondary areas do not overlap. In the next section, we derive a probabilistic guarantee for such a method.

V. Probabilistic Guarantee for Finding the Parent Transformer via Clustering

As discussed in Section I, voltage measurements are generally unavailable on distribution transformers. In this section, we provide proof of a probabilistic performance guarantee for obtaining the mapping between the true meter clusters and the parent transformer using the location information. For example, we use the street categorical information and the XY location information for meters and transformers. The street categorical information filters out meter clusters and transformers belonging to a street. Such a step is valid as it is usually practiced in distribution systems. Next, meter clusters are mapped to the transformers based on the nearness of the XY location.

Assume k transformers supply a radial distribution system. Therefore, there are k meter groups. Let $\mathbf{X}_{j,i}$ be the random variable for the XY coordinates of the j-th parent transformer considering the i-th meter. For instance, around each meter

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

i, $\mathbf{X}_{j,i}$ has a 2-D Gaussian distribution with a peak at the meter itself. Mathematically, the probability density is given as $\mathbf{X}_{j,i} \sim \mathbb{P}_{X_{j,i}}(\mathbf{x}|i, \mathbf{l}^i, j) = \mathcal{N}(\mathbf{l}^i, \sigma^2)$, where $\mathbf{l}^i$ represents the XY coordinates of the i-th meter, and σ represents the standard deviation of the distance between meters and their parent transformers. Let $C(j)$ be the set of meters supplied by the j-th transformer. Hence, the location of the j-th transformer $\mathbf{X}_j$ considering all $C(j)$ meters is the average of $\mathbf{X}_{j,i}$ for all i. For example, $\mathbf{X}_j = \frac{1}{|C(j)|}\sum_{i\in C(j)} \mathbf{X}_{j,i}$, where the modulus sign $|\cdot|$ represents the number of elements in a set. Since all $\mathbf{X}_{j,i}$ are independent, their average is distributed as

$$\mathbf{X}_j \sim \mathcal{N}\left(\frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}, \frac{\sigma^2}{|C(j)|}\right), \tag{5}$$

which suggests that the most likely value of finding the transformer is the mean location of the downstream smart meters. σ can be estimated from the data using Bessel's correction formula. For instance, let a sample consist of n smart meters. Then the distance between a meter i and its parent transformer j is given as $\left\|\mathbf{l}^i - \mathbf{X}_j\right\|_2^2$. Let $\mu := \frac{1}{n}\sum_j \sum_{i\in C(j)} \left\|\mathbf{l}^i - \mathbf{X}_j\right\|_2$ be the mean of the distance between meters and transformers. The estimator of σ is given via Bessel's correction formula as $\sqrt{\frac{1}{n-1}\sum_j \sum_{i\in C(j)} \left(\left\|\mathbf{l}^i - \mathbf{X}_j\right\|_2 - \mu\right)^2}$. Moreover, the greater the number of meters $|C(\cdot)|$ supplied, the lesser the variance will be, which suggests that the transformer is more likely to be found at the mean value of the daughter smart meter locations. We know the probability density function of the k-th transformer location.

It is required to know the probability of failure of the proposed method-based transformer mapping identification. For computing such a probability, we assume transformers supply meters in the streets sequentially. For example, as primary distribution lines extend along the street, the transformers linearly partition the set of meters along the length of the street. The probability of the event so that $\mathbf{X}_j$ is further to $\frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}$ than the next neighboring transformer's location $\mathbf{X}_{j+1}$. Mathematically,

$$\mathbb{P}_X\left(\left\|\mathbf{X}_j - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}\right\|_{L_2} \geq \left\|\mathbf{X}_{j+1} - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}\right\|_{L_2}\right), \tag{6}$$

$$= \mathbb{P}_X\left(\left\|\frac{\mathbf{X}_j - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}}{\sigma\Big/\sqrt{|C(j)|}}\right\|_{L_2} \geq \left\|\frac{\mathbf{X}_{j+1} - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{m_i}}{\sigma\Big/\sqrt{|C(j)|}}\right\|_{L_2}\right),$$

$$= \mathbb{P}_X\left(\left\|\mathbf{Z}_j\right\|_{L_2} \geq \left\|\frac{\mathbf{X}_{j+1} - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}}{\sigma\Big/\sqrt{|C(j)|}}\right\|_{L2}\right).$$

Let $d_{j,j+1}$ define the distance between the mean position of meters for nearby transformers $d_{j,j+1} =: \frac{\sum_{i\in C(j+1)} \mathbf{l}^i}{|C(j+1)|} - \frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}$

$$= \mathbb{P}_X\left(\left\|\mathbf{Z}_j\right\|_{L_2} \geq \left\|\frac{\left(\frac{\sigma \mathbf{Z}_{j+1}}{\sqrt{|C(j+1)|}}\right) + d_{j,j+1}}{\sigma\Big/\sqrt{|C(j)|}}\right\|_{L_2}\right),$$

$$= \mathbb{P}_X\left(\left\|\mathbf{Z}_j\right\|_{L_2} \geq \left\|\frac{\mathbf{Z}_{j+1}\sqrt{|C(j)|}}{\sqrt{|C(j+1)|}} + \frac{d_{j,j+1}}{\sigma\Big/\sqrt{|C(j)|}}\right\|_{L_2}\right).$$

Without loss of generality, let us assume our frame of reference (y-axis) is aligned with the street. Therefore, we consider variation in one axis only.

$$\mathbb{P}_X\left(|Y_j| \geq \left|\frac{Y_{j+1}\sqrt{|C(j)|}}{\sqrt{|C(j+1)|}} + \frac{d_{j,j+1}}{\sigma\Big/\sqrt{|C(j)|}}\right|\right).$$

Since neighboring transformers usually have a similar number of smart meters, we can consider the ratio $\frac{\sqrt{|C(j)|}}{\sqrt{|C(j+1)|}} \approx 1$

$$\mathbb{P}_X\left(|Y_j| \geq \left|Y_{j+1} + \frac{d_{j,j+1}}{\sigma\Big/\sqrt{|C(j)|}}\right|\right),$$

where $d_{j,j+1}$ is the distance between neighboring transformers on the same street, and σ is the average distance between a meter and a transformer. Usually, the distance between two neighboring transformers is at least twice the distance between a meter and a transformer. Therefore, $d_{j,j+1}/\sigma \geq 2$. Furthermore, let us consider the range of values of $|C(j)|$ for all j between c_l and c_h. For instance, if the minimum value $c_l = 4$, the value of $\frac{d_{j,j+1}}{\sigma\Big/\sqrt{|C(j)|}} \geq 4$.

$$\mathbb{P}_X\left(|Y_j| \geq |Y_{j+1} + 4|\right). \tag{7}$$

Given that all Y_i are standard normal variables, the empirical probability computed using 10^6 samples in (7) corresponds to around 0.232%.

Similarly, the probability of failure considering $\mathbf{X}_j$ is further to $\frac{\sum_{i\in C(j)} \mathbf{l}^i}{|C(j)|}$ than the previous neighboring transformer's location $\mathbf{X}_{j-1}$ is given as $\mathbb{P}_X\left(|Y_j| \geq |Y_{j-1} + 4|\right)$, which is the same as (7). Hence, the probability of failure is twice $\mathbb{P}_X\left(|Y_j| \geq |Y_{j+1} + 4|\right)$, which corresponds to around $0.464\% \approx 0.5\%$ for 10^6 samples. Therefore, the empirical probability of success computed using 10^6 samples is $100\% - 0.5\% = 99.5\%$

The first assumption is having the street categorical information so that meters are supplied by transformers in the same street. Using such an assumption, we only consider meters present in the same street. Such an assumption is the usual practice in distribution systems, so it is valid. Moreover, we assume transformers sequentially supply meters in the streets. For example, as primary distribution lines extend along the street, the transformers linearly partition the set of meters along the length of the street.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

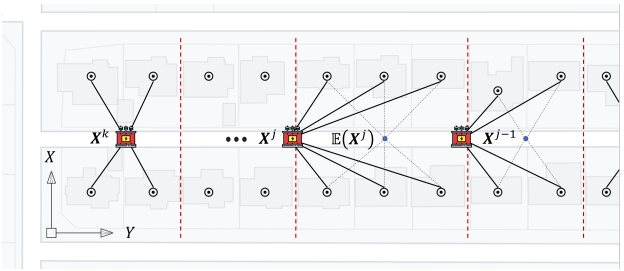


Fig. 3. Partition of meter set along the length of the street by transformers shown by dotted red lines. The primary line is assumed along the street (not shown in the figure).

VI. Meter-Transformer Mapping for Long Streets

In the case of long streets with many transformers, recovering meter-transformer mapping can be a challenge. For instance, due to many transformers, there are many possibilities to assign a meter to a non-parent transformer, while only one possibility for assigning to a parent transformer. Therefore, there is a higher likelihood of making an incorrect decision by the algorithm. Furthermore, such an effect is more profound for higher primary voltage magnitudes since the impedance of the primary lines is small when referred to the secondary, resulting in similar secondary voltage magnitudes for meters close to different transformers. Finally, for more transformers, there may also be similar consumption for two transformer secondary circuits, resulting in similar voltage magnitudes.

To resolve this issue, our algorithm also includes the distance along the street (DAS) metric. The reference for such a distance is one of the meters starting the street. Such information can effectively distinguish between meters supplied by various transformers. As opposed to the usual Haversine distance, the DAS metric does not distinguish between meters on both sides of the street. Moreover, DAS closely aligns with the fact that overhead and underground electric lines follow streets. Finally, the Haversine metric fails on curved streets, however, the DAS effectively identifies the distance covered by electrical lines. Therefore, such metrics aid the voltage magnitude-based algorithm in determining the meter-transformer mapping effectively, even for long streets.

VII. Influence on Other Estimation Problems & Voltage Measurement Errors

A. Influence on Estimation Problems

1) Influence on State Estimation: A thorough analysis of state estimation [37] in a distribution system is given in Reference [38]. The restricted observability (unavailability of nodal admittance matrix) issue is attempted to be solved by a number of contemporary approaches, such as [39], by inserting pseudo measurements into the system. The pseudo measures taken into account include those that use average power values rather than the actual power values, which could not be a true reflection of the measurement. The meter-transformer mapping produced by our approach, however, can aid in providing an exact estimation of the power passing through transformers.

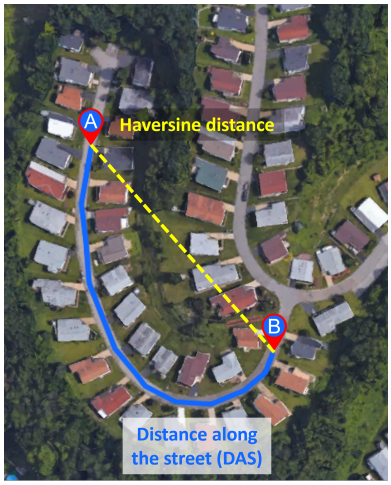


Fig. 4. Comparison between the Haversine distance (dashed yellow line) and distance along the street (solid blue line).

Hence, one can use the obtained power flowing through transformers as measurements. More precise measurements can be used to overcome the problem of restricted observability [25].

2) Influence on Topology Estimation: Topology estimation is a challenging problem that calls for a variety of data. Distribution systems, in contrast to transmission systems, might not include a large range of sensors at several places [5]. To resolve this problem, researchers usually approximate by making ideal assumptions of an isolated subnetwork or the availability of measurement at every system node. Both presumptions might not apply to distribution systems [24]. However, using our method as a preprocessor to identify the meter-transformer mapping accurately, the topology identification methods can use this mapping as an input to identify the remaining topology.

3) Influence on Line Parameter Estimation: To estimate line parameters assuming a pi-model, researchers often assume voltage measurements at both ends of a line [40]. However, measurements are rarely made at both ends of a distribution line, especially in the low-voltage range. Therefore, our method can be used as a preprocessing step by identification of the meter-transformer mapping. Such a mapping can help the line parameter estimation. Assume a pad-mounted transformer is used to power 14 smart meters. It is because, for pad-mounted transformers, there is usually a dedicated cable from the meter to the transformer for easy replacement of the cable when needed. So our method has significance for line parameter estimation.

B. Impact of Voltage Measurement Errors on Estimation Problems

1) Common Mode Error: Common mode errors can arise when the ground terminals of the measuring device and the voltage source are not at the same potential [41]. Since the ground terminal of a smart meter is the same as the load, a smart meter may not experience this issue.

2) AC Loading Error: The high source impedance and low internal resistance of the meter lead to AC loading error [42].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

However, since the electric distribution network supplies high electric power, it has a very low source resistance. As a result, this inaccuracy barely affects our mapping method and the estimating techniques.

3) Crest Factor Error: The AC waveforms that a smart meter measures are not true sine waveforms, e.g., the non-linear effect of transformer core clips the peaks of the waveform. Meters often calculate the RMS voltage by multiplying the average voltage of the full-wave rectified AC signal by the ratio $\left(\frac{1}{\sqrt{2}}\right) \times \left(\frac{\pi}{2}\right)$ leading to crest factor errors. However, the voltage non-linearity caused by the transformer will be the same for all the meters supplied by the transformer. Therefore, even though crest factor error may lead to inaccurate RMS measurement, it has little impact on our mapping.

4) Quantization Error: Quantization error results due to the low resolution of the analog-to-digital converter (ADC) for measurement [43]. For example, in our case, the smart meter ADC has a resolution of around 0.025 V. However, such an error can be neglected given the allowed voltage range, i.e., $\pm 5\%$ of the rated voltage. For example, if the rated voltage is 120 V, the allowed voltage variation is around 12 V. Therefore, the estimation method using our mapping method will remain unaffected.

5) Smart Meter Analog-to-Digital Converter (ADC) Malfunctioning: In such a case, the estimating methods that employ our mapping method will be ineffective. However, the smart meter voltage reading may not stay within the permitted range in this scenario, which is $\pm 5\%$ of the rated voltage. Therefore, the relevant electric distribution company (EDC) can identify and replace such a meter.

There are many methods that can be used as a preprocessing step before our algorithm to detect smart meter malfunction, [44], [45]. In case of a smart meter malfunction, our method can work by excluding the malfunctioning meters, and recovering the topology for the working meters.

VIII. Operating Conditions

A. Power Loads

Power load means the loads that mainly consume active power, e.g., heaters and boilers [46]. Power loads use a lot of active power and a little reactive power Due to parasitic inductance and capacitance. When power loads are turned on, their energy use remains largely constant. Power loads are switched on for longer periods of time than other loads, which turn on briefly.

At first glance, it could appear that power loads use power pretty consistently, making it difficult to detect their use. However, the thermostat turns them off and on to keep the desired temperature. Such a signature is also reflected in the voltage profile that is recorded in the voltage data input to the algorithm. As a result, our method can effectively handle the power load.

B. Renewables

The power output profiles are similar for renewable energy sources like solar PV. They are typically connected behind the meter. However, because of their similar profiles, the meters with PV supplying behind the meter may have similar voltages [47]. For identification techniques that solely consider voltage data, this might be an issue.

However, this paper also focuses on using location data intelligently. For example, we divide the smart meter data into sub-street clusters using innovative categorical location information, such as street names, zip codes, etc., in addition to the voltage data. We propose employing the innovative distance along the street (DAS) measure to determine the meter-transformer mapping once the smart meter data has been divided into streets. Therefore, our method is robust.

C. Switching

In the case of connection switching, e.g., a meter switch from one transformer to another transformer, we use change-point detection [48] to detect the timestep corresponding to such changes. We separate the datasets before and after the change-point to recover the two meter-transformer mappings separately.

IX. Numerical Validation

A. Data Description

The simulations are implemented on the IEEE Power and Energy Society (PES) distribution networks for IEEE benchmark systems, such as 123-bus systems. In addition, our method is deployed on a high penetration medium voltage (MV) feeder from a southwestern utility in the United States. The feeder contains around 2,600 smart meters (including solar meters), 1,737 customers, and 371 service transformers. In addition, the feeder contains 4 MV capacitor banks, 765 PV modules, and a total of 2,283 switches in the system. The total installed PV capacity is around 765 kW. The length of underground cables in the system is 130,956 meters.

B. Validation strategy

A detailed OpenDSS model of the feeder is available from the utility as part of a recent project from the United States Advanced Research Projects Agency-Energy (ARPA-E). The meter-transformer mapping ground truth is extracted from the model and used to validate our algorithm.

The computational environment consists Intel(R) Core(TM) i7-10510U CPU with 16 GB RAM. We used Python with Anaconda Spyder to code, debug, and evaluate the algorithm on testbench systems and real utility systems.

C. Validation of the Proposed Method on Benchmark Systems

In order to show that the accuracy of our proposed method is independent of the benchmark system, the validation is shown on three benchmark systems, i.e., the IEEE 8-bus system, the IEEE 69-bus radial distribution system from Das, and the IEEE-123 bus system.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

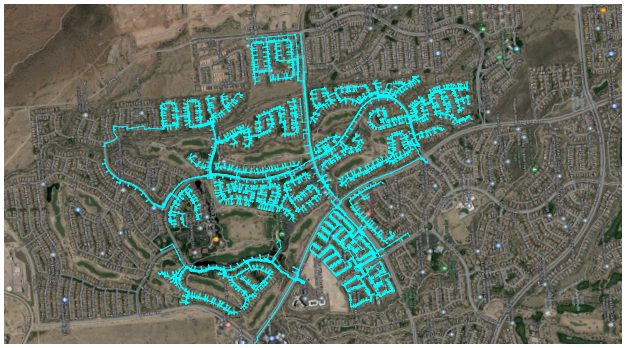


Fig. 5. Big picture of the high penetration feeder from our partner utility.

1) Validation using the IEEE 8-bus system: We begin with a simple analysis consisting of an 8-bus system. In the vast majority of relay coordination literature, the 8-bus system is frequently utilized as a standard test case [50]. We divide the system into two streets with one transformer in each street to construct the test case, as shown in Fig. 7. We also separate the street locations. The first step of our algorithm accurately separates the two streets, identifying the meter-transformer mapping directly. The execution time for the first step of our algorithm is 21 ms.

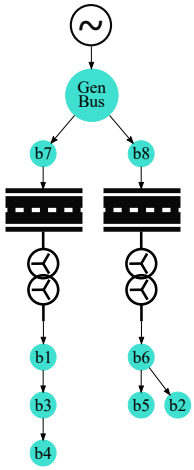


Fig. 7. IEEE 8-bus system modified and divided into two streets with one transformer in each street.

2) Validation using the IEEE 69-bus system: The IEEE 69-bus radial distribution system from [51] has been modified into two streets, as shown in Fig. 9. The first street contains only one transformer, while the second street contains two transformers. The goal is to obtain the meter-transformer mapping. The transformers are considered step-down transformers. Hence, the conductors' impedances between the two transformers are negligible when referred to their secondary sides. Therefore, if the net consumption of the loads belonging to different transformers is similar, the voltage magnitudes of the meters nearer to two different transformers will be similar. Thus, they may be grouped together rather than being grouped with their sibling meters. However, using street categorical information aids the voltage magnitude-based clustering to segment street-wise. Furthermore, within each street, the distance along the street (DAS) metric aids the voltage magnitude-based clustering to obtain the correct meter-transformer mapping.

3) Validation on the IEEE 123-bus system: To validate the chosen algorithm on a bigger system, the IEEE 123-bus system was selected and modified, as shown in Fig. 10. Similar to the IEEE 69-bus system, we created two streets with three transformers each. To make the system more complex, we assume the same node feeding all transformers of a street, as shown. Such an assumption will make the voltage magnitudes of the transformer secondary nodes more similar to each other than otherwise. Hence, using voltage magnitudes alone does not recover the correct meter-transformer mapping no matter the chosen algorithm. However, using the proposed method identifies the true meter-transformer mapping.

TABLE I
COMPARISON OF APPROACHES AND METHODS ON THE MODIFIED IEEE-123 BUS SYSTEM.

Method	AMI Score	Adjusted Rand Score	Execution Time
The Proposed Method	1.000	1.000	1.84 s
Spectral Clustering-Voltage	0.509	0.356	0.13 s
Kmeans-Voltage	0.618	0.503	0.04 s
BIRCH-Voltage	0.660	0.468	0.36 s
DBSCAN-Voltage	0.598	0.535	0.08 s

Voltage refers to voltage magnitudes only.
AMI refers to Adjusted Mutual Information.

D. Validation of the Proposed Method on Real Systems

In addition to the benchmark systems, we also show the capability of our method on real systems. For instance, we show how we segment the data into street clusters using voltage magnitude and street categorical information and how we further separate the street clusters by transformers to obtain transformer clusters. Finally, we show a large-scale example of the recovered meter-transformer mapping from the complete feeder of the partner utility.

1) Validation of Segmentation of Data into Streets: Fig. 11 shows the separation of utility systems into streets, which is needed to use the distance along the street information. Such a segmentation uses street categorical information to aid voltage magnitude information via density-based clustering.

2) Validation of Streetwise Meter-Transformer Mapping Identification: Once the data is separated into streets, the next step is to obtain the distance along the street information for each house. Our chosen reference point is the street end with the lowest block number. Next, we obtain the DAS using an online map service, e.g., Bing Maps API. Finally, we use DAS with the voltage magnitude information to obtain the meter-transformer mapping, as shown in Fig. 6c.

E. Bulk Area Validation

We performed extensive validation of the proposed method on the entire feeder from our partner utility. Fig. 12 shows the results of the bulk area validation we performed on the utility area. For example, we can see that the proposed algorithm correctly identified the meter-transformer mapping.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

(a) The recent method [49] for meter-transformer mapping identified only 3 clusters, which is incorrect.

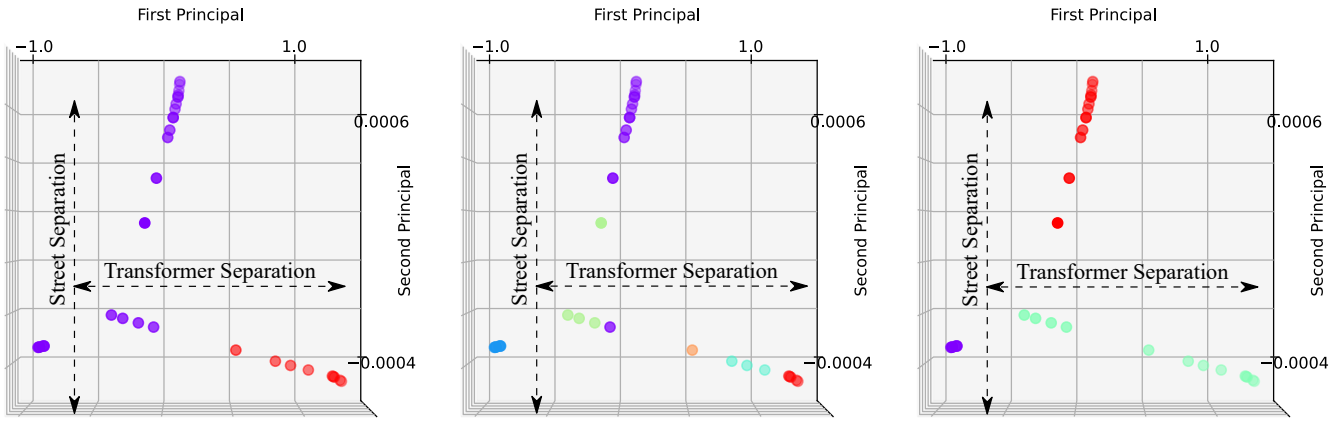
(b) Regular spectral clustering with voltage magnitude only does not work well in scenarios with high PV penetration due to voltage magnitude similarity. Erroneous results are marked with red arrows.

(c) Distance along the street (DAS) metric guides the algorithm on how the distribution line extends from one section to another, following the street. The corrected results are marked with green arrows.

Fig. 6. Comparison of a recent method [49], spectral clustering without using the DAS metric, and spectral clustering using the DAS metric.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581



(a) Result of the $k-$means clustering algorithm. Specify $k = 2$ for $k-$means, but it identified incorrect clusters since the within-cluster optimization approach is unsuitable for voltage magnitude data. Execution time for $k-$means is 44 ms.

(b) Result of the BIRCH clustering algorithm. BIRCH forms three instead of two clusters. BIRCH is not suitable for voltage magnitude data since the radius hyperparameter needs to be hard coded. Execution time for BIRCH is 39 ms.

(c) Result of spectral clustering algorithm with voltage magnitude mutual information. Spectral clustering gave correct results for the transformer separation using voltage magnitude data alone. Execution time for spectral clustering is 22 ms.

Fig. 8. Comparison of the three clustering algorithms using both voltage magnitude and location data on the modified IEEE-69 bus test feeder with two streets and three transformers, as shown in Fig. 9. The dotted line for "Transformer Separation" shows the ground truth of the two clusters.

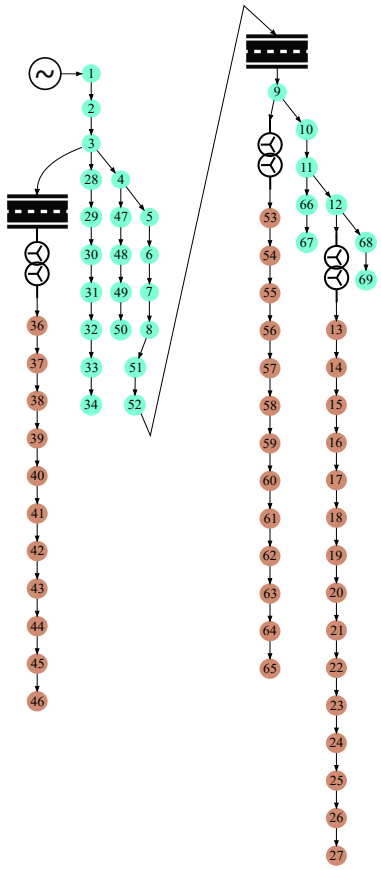


Fig. 9. A realistic scenario showing the IEEE-69 bus feeder with two streets and three transformers. The secondary nodes contain smart meters.

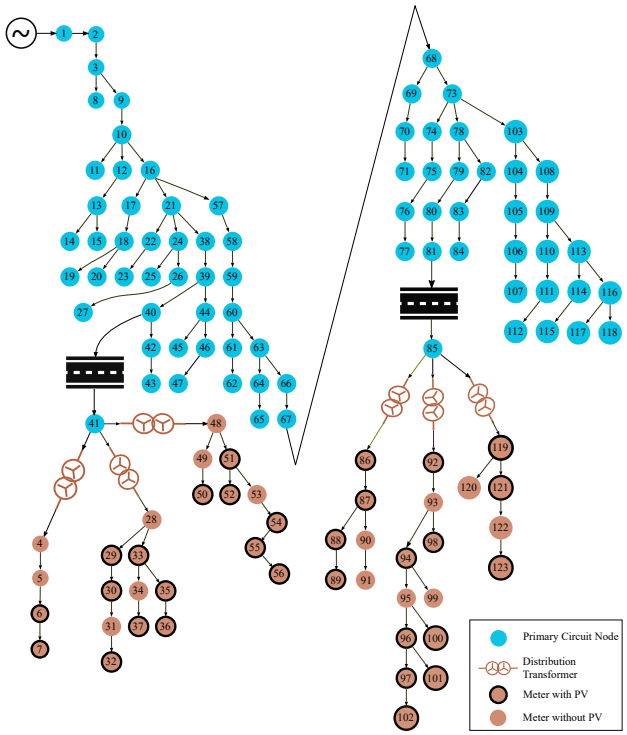


Fig. 10. A realistic scenario using the IEEE-123 bus system, which considers negligible impedance of the primary circuit conductor between the transformers of a street.

F. Validation of Guarantee Probability

Using the data of the real feeder given above, we validate the probability proved in Section V, which suggests at least 99.5% accuracy for mapping transformers to meter clustering using the proposed method. We select a sample of 322 transformers and their secondary circuits comprising 1,584 smart meters. Here, we validate the assumption that the transformer nearest to the true meter cluster is the parent transformer of the cluster with at least 99.5% accuracy. Below is the bar chart showing the correct results. For example, we get 100% probability based on the sample.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

Fig. 11. The street categorical information helps voltage magnitude to identify meters related to the street clusters accurately.

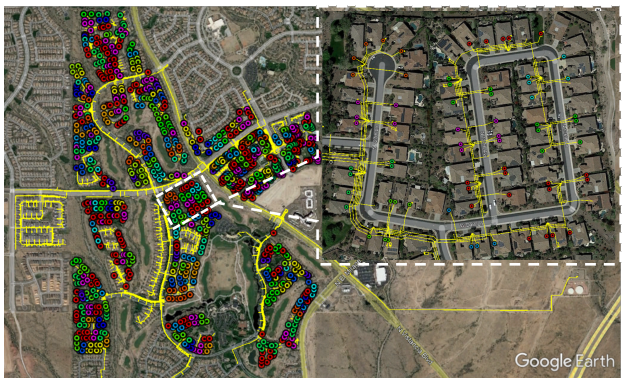


Fig. 12. Accurate results from our proposed method on the test system of the high penetration PV feeder available from our partner utility. The node colors indicate the meter clusters, whereas the yellow lines indicate the ground truth circuit diagram.

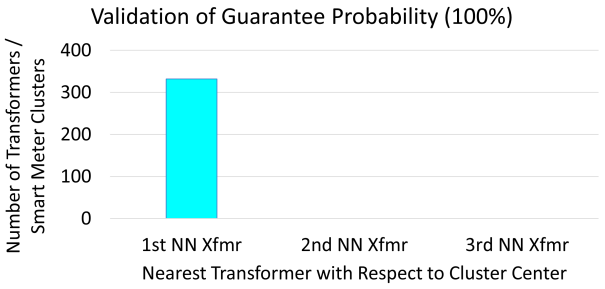


Fig. 13. Validation of the assumption that the transformer nearest to the true meter cluster is the parent transformer of the cluster with at least 99.5% accuracy.

G. Impact of Voltage Measurement Accuracy

In section VII, we discussed the various types of voltage measurement errors and the impact they can have on our method and other estimation methods. For instance, we try to implement two recent papers on topology identification [52] and [53]. We show numerical validation and comparison of our method against the two recent methods. For example, we can observe that our method happens to be more robust than the two recent methods.

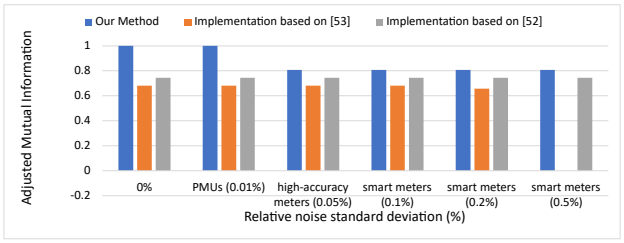


Fig. 14. The bar plot shows the adjusted mutual information between the ground truth topology and the algorithmic outcome. For example, it shows that our method gives 100% correct results if the measurement errors are lesser than or equal to 0.01%.

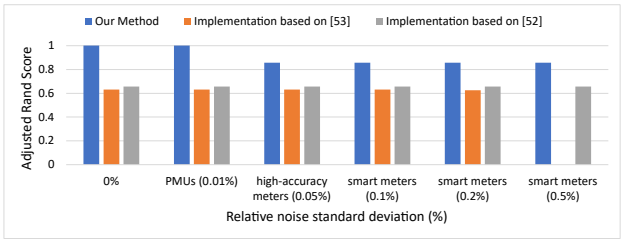


Fig. 15. The bar plot shows the adjusted Rand score between the ground truth topology and the algorithmic outcome. For instance, it shows that our method gives 100% correct results if the measurement errors are lesser than or equal to 0.01%.

Based on the validations in Sections IX-C, IX-D, IX-E, and IX-G, our proposed algorithm correctly identified the meter-transformer mapping even for challenging scenarios.

X. Conclusion

While distributed energy resources (DERs) and electric vehicles (EVs) introduce benefits, they increase the power through a distribution transformer, requiring meter-transformer mapping. Previous methods mostly ignored the location information that is widely available. This paper focuses on using the location information in a better way to aid voltage magnitude-based identification. The first step involves street categorical information that intelligently separates the data streetwise, and the second step uses the distance along the street information as a feature. By constructing complicated scenarios using the IEEE 69-bus and IEEE 123-bus systems, we compared and validated our algorithm, whereas previously known methods could not obtain 100% accuracy. Future work can identify the probabilistic guarantee to recover the meter-transformer mapping using both voltage magnitude and location information.

References

[1] "Ubigrid DTM | Ubicquia," https://www.ubicquia.com/products/distribution-transformer-monitoring-DTM, [Online; accessed 2022-12-21].
[2] L. M. Adesina, A. Abdulkareem, O. Ogunbiyi, and O. Ibrahim, "On-load measurement method for the reliability of distribution transformers," *MethodsX*, vol. 7, no. 101089, 2020.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

[3] Y. Weng, A. Kumar, M. B. Saleem, and B. Zhang, "Big data and deep learning platform for terabyte-scale renewable datasets," *IEEE Power Systems Computation Conference*, 2018.

[4] B. Saleem, Y. Weng, and F. M. Gonzales, "Association rule mining for localizing solar power in different distribution grid feeders," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2589–2600, 2020.

[5] Y. Weng, R. Negi, and M. D. Ilić, "A search method for obtaining initial guesses for smart grid state estimation," in *International Conference on Smart Grid Communications*, 2012.

[6] G. N. Korres and N. M. Manousakis, "A state estimation algorithm for monitoring topology changes in distribution systems," *IEEE Power and Energy Society General Meeting*, 2012.

[7] J. Huang, V. Gupta, and Y.-F. Huang, "Electric grid state estimators for distribution systems with microgrids," *Annual Conference on Information Sciences and Systems*, 2012.

[8] G. Zhang, S. Lee, R. Carroll, J. Zuo, L. Beard, and Y. Liu, "Wide area power system visualization using real-time synchrophasor measurements," *IEEE Power and Energy Society General Meeting*, 2010.

[9] D. Deka, M. Chertkov, and S. Backhaus, "Topology estimation using graphical models in multi-phase power distribution grids," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1663–1673, 2019.

[10] Y. Weng, M. D. Ilić, Q. Li, and R. Negi, "Distributed algorithms for convexified bad data and topology error detection and identification problems," *International Journal of Electrical Power & Energy Systems*, vol. 83, pp. 241–250, 2016.

[11] Y. Weng, M. D. Ilić, Q. Li, and R. Negi, "Convexification of bad data and topology error detection and identification problems in ac electric power systems," *IET Generation, Transmission and Distribution*, vol. 9, no. 16, pp. 2760–2767, 2015.

[12] J. Yu, Y. Weng, and R. Rajagopal, "Data-driven joint topology and line parameter estimation for renewable integration," *IEEE Power and Energy Society General Meeting*, 2017.

[13] Y. Weng, C. Faloutsos, and M. D. Ilić, "Data-driven topology estimation," *IEEE International Conference on Smart Grid Communications*, 2014.

[14] G. Cavraro, R. Arghandeh, G. Barchi, and A. von Meier, "Distribution network topology detection with time-series measurements," *IEEE International Innovative Smart Grid Technologies Conference*, 2015.

[15] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," *Conference on Decision and Control*, 2013.

[16] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2682–2694, 2016.

[17] J. Yu, Y. Weng, and R. Rajagopal, "PaToPaEM: A data-driven parameter and topology joint estimation framework for time-varying system in distribution grids," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1682–1692, 2018.

[18] J. Yu, Y. Weng, and R. Rajagopal, "PaToPa: A data-driven parameter and topology joint estimation framework in distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4335–4347, 2017.

[19] H. Li, Y. Weng, Y. Liao, B. Keel, and K. E. Brown, "Distribution grid impedance & topology estimation with limited or no micro-PMUs," *International Journal of Electrical Power & Energy Systems*, vol. 129, no. 106794, pp. 1–12, 2021.

[20] J. Zhang, Y. Wang, Y. Weng, and N. Zhang, "Topology identification and line parameter estimation for non-PMU distribution network: A numerical method," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4440–4453, 2020.

[21] Y. Liao *et al.*, "Unbalanced multi-phase distribution grid topology estimation and bus phase identification," *IET Smart Grid*, 2019.

[22] C. Rudin *et al.*, "Machine learning for the new york city power grid," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 328–345, 2011.

[23] D. Deka, S. Backhaus, and M. Chertkov, "Structure learning in power distribution networks," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1061–1074, 2017.

[24] B. Saleem, Y. Weng, E. Cook, H. Wang, E. Blasch, and D. Peachera, "Recover meter-transformer connectivities using geospatial unsupervised learning and Q-GIS," in *2022 IEEE Power & Energy Society General Meeting*, 2022.

[25] B. Saleem and Y. Weng, "Explainable graph theory-based identification of meter-transformer mapping," 2022. [Online]. Available: https://arxiv.org/abs/2205.09874

[26] J. D. Watson, J. Welch, and N. R. Watson, "Use of smart-meter data to determine distribution system topology," *The Journal of Engineering*, vol. 2016, no. 5, pp. 94–101, 2016.

[27] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964–1971, 2015.

[28] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996.

[29] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[30] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[31] H. Mehta, P. Kanani, and P. Lande, "Google maps," *International Journal of Computer Applications*, vol. 178, no. 8, pp. 41–46, 2019.

[32] R. Al Abri, E. F. El-Saadany, and Y. M. Atwa, "Optimal placement and sizing method to improve the voltage stability margin in a distribution system using distributed generation," *IEEE transactions on power systems*, vol. 28, no. 1, pp. 326–334, 2012.

[33] T. S. Basso and R. D. Deblasio, "Ieee p1547-series of standards for interconnection," in *2003 IEEE PES Transmission and Distribution Conference and Exposition (IEEE Cat. No. 03CH37495)*, vol. 2. IEEE, 2003, pp. 556–561.

[34] R. Walling, R. Saint, R. C. Dugan, J. Burke, and L. A. Kojovic, "Summary of distributed resources impact on power delivery systems," *IEEE Transactions on power delivery*, vol. 23, no. 3, pp. 1636–1644, 2008.

[35] M. I. Jordan and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, pp. 849–856, 2002.

[36] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," *Knowledge Discovery and Data Mining*, vol. 96, no. 34, pp. 226–231, 1996.

[37] W. Stevenson and J. Grainger, *Power System Analysis*. McGraw-Hill Education, 1994. [Online]. Available: https://books.google.com/books?id=NBIoAQAAMAAJ

[38] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3875–3883, 2016.

[39] S. Naka, T. Genji, T. Yura, and Y. Fukuyama, "A hybrid particle swarm optimization for distribution state estimation," *IEEE Transactions on Power systems*, vol. 18, no. 1, pp. 60–68, 2003.

[40] Y. Liao and M. Kezunovic, "Online optimal transmission line parameter estimation for relaying applications," *IEEE Transactions on Power Delivery*, vol. 24, no. 1, pp. 96–102, 2008.

[41] G. Edwards and I. Watson, "A study of common-mode failures," UKAEA Risley Nuclear Power Development Establishment, Tech. Rep., 1979.

[42] N. M. Vrana, "Compensation for instrument losses in circuits containing a wattmeter, voltmeter, and ammeter," *Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics*, vol. 76, no. 3, pp. 325–327, 1957.

[43] X. Lan, Q. Hu, and J. Cheng, "Revisting quantization error in face alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1521–1530.

[44] Q. Zhao, J. Mu, X. Han, D. Liang, and X. Wang, "Evaluation model of operation state based on deep learning for smart meter," *Energies*, vol. 14, no. 15, p. 4674, 2021.

[45] M. Liu, D. Liu, G. Sun, Y. Zhao, D. Wang, F. Liu, X. Fang, Q. He, and D. Xu, "Deep learning detection of inaccurate smart electricity meters: a case study," *IEEE Industrial Electronics Magazine*, vol. 14, no. 4, pp. 79–90, 2020.

[46] F. Ygge and J. M. Akkermans, *Power load management as a computational market*. Högskolan i Karlskrona/Ronneby, 1996.

[47] J. Chen, X. Xu, Z. Yan, and H. Wang, "Data-driven distribution network topology identification considering correlated generation power of distributed energy resource," *Frontiers in Energy*, vol. 16, no. 1, pp. 121–129, 2022.

[48] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[49] K. Ashok, D. Divan, and F. Lambert, "Grid edge analytics platform with AMI data," in *IEEE Power and Energy Society Innovative Smart Grid Technologies Conference*, 2018.

[50] H. Zeineldin, E. El-Saadany, and M. Salama, "Optimal coordination of overcurrent relays using a modified particle swarm optimization," *Electric Power Systems Research*, vol. 76, no. 11, pp. 988–995, 2006.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3281581

[51] W. Gu, L. Luo, T. Ding, X. Meng, and W. Sheng, "An affine arithmetic-based algorithm for radial distribution system power flow with uncertainties," *International Journal of Electrical Power & Energy Systems*, vol. 58, pp. 242–245, 2014.
[52] G. Cavraro and A. Bernstein, "Bus clustering for distribution grid topology identification," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4080–4089, 2020.
[53] V. L. Srinivas and J. Wu, "Topology and parameter identification of distribution network using smart meter and μpmu measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

Bilal Saleem (Student Member, IEEE) received the B. Tech. degree in electrical engineering from the National University of Science and Technology, Pakistan; and M.Sc. degree in electrical engineering from Politecnico di Milano, Italy. He received his Ph.D. degree from Arizona State University, Tempe, AZ, USA.

Dr. Saleem was awarded the University Graduate Fellowship by Arizona State University in 2018. He also received Thesis Abroad Scholarship from Politecnico di Milano, Italy, in 2017. In addition, he received fully funded scholarship for M.Sc. in Italy. He secured the second position at the IEEE-Technovate project exhibition competition. He also received Merit certificates and cash prizes for securing high GPA National University of Science and Technology, Pakistan. His current research interest is in the interdisciplinary area of AI applications in power systems. In particular, his research focuses on topology identification in distribution grids using AMI data and GIS information.

Yang Weng (Senior Member, IEEE) received the B.E. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China; the M.Sc. degree in statistics from the University of Illinois at Chicago, Chicago, IL, USA; and the M.Sc. degree in machine learning of computer science and M.E. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, USA. After finishing his Ph.D., he joined Stanford University, Stanford, CA, USA, as the TomKat Fellow for Sustainable Energy. He is currently an Assistant Professor of electrical, computer and energy engineering at Arizona State University (ASU), Tempe, AZ, USA. His research interest is in the interdisciplinary area of power systems, machine learning, and renewable integration.

Dr. Weng received the CMU Dean's Graduate Fellowship in 2010, the Best Paper Award at the International Conference on Smart Grid Communication (SGC) in 2012, the first ranking paper of SGC in 2013, Best Papers at the Power and Energy Society General Meeting in 2014, ABB fellowship in 2014, Golden Best Paper Award at the International Conference on Probabilistic Methods Applied to Power Systems in 2016, and Best Paper Award at IEEE Conference on Energy Internet and Energy system Integration in 2017, Best Paper Award at the IEEE North American Power Symposium in 2019, and Best Paper Award at the IEEE Sustainable Power and Energy Conference in 2019.

Vijay Vittal (Life Fellow, IEEE) received the Ph.D. degree from Iowa State University, Ames, in 1982. He is a Regents' Professor and the Ira A. Fulton Chair Professor in the School of Electrical, Computer and Energy Engineering at Arizona State University, Tempe. Dr. Vittal is a member of the National Academy of Engineering.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/