Electrical Power and Energy Systems 142 (2022) 108291

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Density-based clustering algorithm for associating transformers with smart meters via GPS-AMI data

Elizabeth Cook [a,1], Muhammad Bilal Saleem [b,1], Yang Weng [b,*], Stephen Abate [c], Katrina Kelly-Pitou [c], Brandon Grainger [c]

[a] Duquesne Light Company, Pittsburgh, 15219, PA, USA
[b] ECEE Arizona State University, Tempe, 85281, AZ, USA
[c] University of Pittsburgh, Pittsburgh, 15260, PA, USA

ARTICLE INFO

Keywords:
Topology clustering
Meter-transformer mapping
Density-based clustering

ABSTRACT

The ongoing deployment of Distributed Energy Resources, while bringing benefits, introduces significant challenges to the electric utility industry, especially in the distribution grid. These challenges call for closer monitoring through state estimation, where real-time topology recovery is the basis for accurate modeling. Previous methods either ignore geographical information, which is important in connectivity identification or are based on an ideal assumption of an isolated sub-network for topology recovery, e.g., within one transformer. This requires field engineers to identify the association, which is costly and may contain errors. To solve these problems, we propose a density-based topology clustering method that leverages both voltage domain data and the geographical space information to segment datasets from a large utility customer pool, after which other topology reconstruction methods can carry over. Specifically, we show how to use voltage and GPS information to infer associations within one transformer area, i.e., to identify the meter-transformer connectivity. To give a guarantee, we show a theoretic bound for our clustering method, providing the ability to explain the performance of the machine learning method. The proposed algorithm has been validated by IEEE test systems and Duquesne Light Company in Pittsburgh, showing outstanding performance. A utility implementation is also demonstrated.

1. Introduction

The power distribution grid has been designed with the idea of one-way power flows from feeders to end-users [1]. However, increasing demand for renewable energy (i.e., photovoltaic and storage devices) changes one-way power flows into two-way power flows [2]. As a large portion of the infrastructure pre-dates modern communication methods, many distribution assets were not designed for two-way power flows. Therefore, Electric Distribution Companies (EDCs) need to have visibility of these assets to avoid potential risks of two-way power flows, e.g., outages and equipment damages. For example, EDCs can use topology information to monitor the power grid in real-time and run associated analyses or optimizations. But, a significant challenge for EDCs to gain such visibility today is a lack of system-wide models of their distribution systems based on accurate topology [3,4].

For identifying topology, one idea is to replicate the approaches used on the transmission grid. But, transmission grid approaches rely on mature telemetry and relatively infrequent topology changes [5–9], making them unsuitable for distribution grid topology reconstruction. For instance, most transmission changes are planned for weeks, months, or even years ahead, allowing the topology model to be updated with high accuracy [10]. Unfortunately, the telemetry infrastructure is limited on the distribution system, which has regular and "unexpected" topology changes [11–13], due to routine but unreported reconfiguration [11,14–16]. Even worse, with the introduction of DER, the distribution grid becomes less predictable with the addition of the intermittent generation and is undergoing multiple reconfigurations and upgrades almost every day of operation for many utilities [17–19].

Fortunately, with the deployment of Advanced Meter Infrastructure (AMI), there is hope for utilities to conduct data mining of the big AMI data set. Therefore, there has been work done to develop methods to use the AMI and micro-synchrophasor data to recover the system topology leveraging voltage correlations [11–13]. For example, [20] estimates system topology using voltage magnitudes. However, it assumes all lines to have the same per unit length inductance to resistance

* Corresponding author.
E-mail address: yang.weng@asu.edu (Y. Weng).
[1] Authors contributed equally.

https://doi.org/10.1016/j.ijepes.2022.108291
Received 6 August 2021; Received in revised form 3 April 2022; Accepted 23 April 2022
Available online 24 May 2022

ratios. [21] uses Chow-Liu algorithm for identification of topology. Additionally, [22] estimates the topology and the line parameters combined using historical real and reactive power, voltage magnitudes, and voltage angle measurements. A common issue of these methods is that they need voltage magnitudes at all the nodes of the system, including poles and transformers. Such assumptions are invalid due to the vast spread of distribution systems, making voltage sensing unavailable at every pole and transformer.

Additionally, existing approaches usually require a tremendous effort to narrow down data so that one can obtain smart meter data within each feeder. However, since the topology is unknown in general, it is hard for a utility to tell which smart meters belong to which transformers beforehand. This difficulty prevents the application of the data-driven approaches above. For example, a pull of data from a medium-size utility can lead to an analysis of voltages from 100,000 customers from different feeders. Therefore, if the information on the parent transformer was unavailable at the time of installation, it is extremely hard to require the utility metering department to assign a feeder label to meters within each feeder. Even worse, such a data pull from smart meters can cover areas where no data is available, e.g., prepaid meters such as M-Power meters or un-metered accounts. A voltage-data-based topology recovery in a large area, e.g., 100,000 customers, will cause incorrect inter-area topology connections covering large distances, especially when a non-metered area is between two metering areas.

So, a better idea is to use GPS information to help analysis purely based on voltage domain. This is because geographically different and unconnected smart meters can have similar voltage profiles leading to errors for algorithms that consider voltage only. Therefore, [23] estimates the voltage at the point of coupling (VPC), which is the point where the customer service wire meets the secondary line. [23] uses VPC and the prior connection information from utility GIS system information for joint decision. But, no theoretical guarantee of correction is provided. In addition, such ideas require a fairly accurate prior knowledge of meter-transformer connection information [23,24], which is hard to obtain in reality.

To resolve these issues, we propose to use a carefully selected and customized clustering method based on both the voltage space and the geographical space. To achieve this goal, we analyze the typical design mechanisms of clustering methods, based on within-cluster distance (e.g., K-means [25]) and the number of points within each cluster (e.g., BIRCH method [26]), and density-based method (e.g., DBSCAN [27]). But, the within-cluster approach is not good when a feeder can be quite long, and the number of points within a cluster can vary among different feeders and utilities. But, a density-based method can work well among different feeders with some adjustment because the density is relatively high along the feeder, surrounded by buildings.

To include the voltage information into the clustering, we also need to quantify the distances in the voltage domain. While there are different metrics, mutual information has been proved to be quite useful in the distribution system analysis via the Chow-Liu algorithm [21]. Therefore, we propose to use mutual information to create a relative distance in the voltage domain and combine it with spatial domain information for density-based clustering. Moreover, we provide a theoretic guarantee for the robustness of our clustering method, giving explainability to the performance of the machine learning method. In particular, we show that adding new data does not break, disrupt, or merge the original clusters. Finally, we show how to improve the method further for robustness.

To summarize, the main contributions of this paper are as follows:

1. We provide a comparative analysis of various topology clustering approaches for power system data. We combine voltage and geographic information in a natural way using density-based clustering.
2. We provide an explainable and intuitive theoretical guarantee for the robustness of our proposed clustering method.
3. We improve our method further for robustness.

Numerical experiments are carried out on the standard distribution testbeds, e.g., IEEE 123-bus, and by our partner EDC's local grid with 10,000 customers. The result shows that the proposed method segments the distribution grids accurately and helps to achieve a highly accurate topology estimate.

The rest of the paper is organized as follows: Section 2 defines topology clustering. Section 3 introduces the proposed clustering methods for integrating utility GPS and public GIS information. Section 4 provides a robustness guarantee. Section 5 validates the idea numerically, and Section 6 concludes the paper.

2. System model

In order to define the problem better, we need to define the graphical backbone of the targeted network clearly. A distribution system is graphically characterized by nodes (buses) $\mathcal{V} = 1, 2, \dots, M$ and by branches $\mathcal{E} = (i, i'), i, i' \in \mathcal{V}$. N leaf nodes in $\mathcal{V}$ are the smart meters, k nodes in $\mathcal{V}$ are the service transformers. The meter-transformer connections are modeled as branches in $\mathcal{E}$. To define the method for topology clustering, we describe time-series voltage data given by smart meters. The voltage at meter i and time t can be represented as $v_i(t) = |v_i(t)| \exp^{j \cdot \theta_i(t)}$, where $|v_i(t)| \in \mathbb{R}$ denotes the magnitude of the bus instantaneous voltage in per-unit, and $\theta_i(t)$ denotes the phase angle of the voltage in radian. The root mean square (RMS) voltage measurements are sampled every 5 mins from meter i to form a vector $\mathbf{v}^i$. The voltage time-series with T timeslots for N smart meters $\mathbf{v}^1, \dots, \mathbf{v}^N \in \mathbb{R}^{T \times 1}$ are stored as row vectors in matrix $\mathbf{V} \in \mathbb{R}^{N \times T}$. In addition to time series data definition, it is equally important to understand how the location data is going to impact our learning. The latitude–longitude pairs in radians for N smart meters $\mathbf{l}^1, \dots, \mathbf{l}^N \in \mathbb{R}^{2 \times 1}$ are stored as row vectors in matrix $\mathbf{L} \in \mathbb{R}^{N \times 2}$. $\mathbb{R}$ represents the set of real numbers. With spatial structure and temporal slots well defined, we will need to quantify the measurements based on them before mathematically defining our problem. For example, the combined dataset becomes $[\mathbf{L}, \mathbf{V}] \in \mathbb{R}^{N \times (T+2)}$ with row vectors $\mathbf{x}^1, \dots, \mathbf{x}^N$ for N smart meters. The k service transformer secondaries form a k-way partition of N smart meters in the distribution grid. $Cluster(j)$ represents a vector of all meters indices supplied by transformer j. Due to radial configuration, a smart meter $i \in \{1, \dots, N\}$ is uniquely present in a cluster $j \in \{1, \dots, k\}$ that is supplied by a common transformer. There exists a many-to-one mapping $f : i \to j$. We define the problem below.

- Problem: Identify smart clustered meters to different transformers
- Given: Smart meter voltage data and location $[\mathbf{V}, \mathbf{L}]$,
- Find: The mapping rule $f : i \to j$.

3. Clustering methods for grid segmentation

3.1. Data preparation

In the past, most topology-related studies in the distribution grid assume to use AMI temporal data only, e.g., voltages. The past methods did not use the location information, although the location information is equally important, and many utilities have such information for usage. Even if a utility does not have the locations of smart meters, utilities can convert building addresses into latitudes and longitudes of the smart meters by using Google Maps API. Similarly, for poles location, a utility can employ a person to obtain fairly accurate GPS coordinates of poles using Google Street View without field visits.

For spatial data preparation, we had the GPS coordinates of transformers, poles, and smart meters. The transformers and poles GPS measurements are usually conducted using accurate GPS measurements, while the GPS inside smart meters is inaccurate. Therefore, we discarded the GPS measurements from meters. Instead, we geocoded

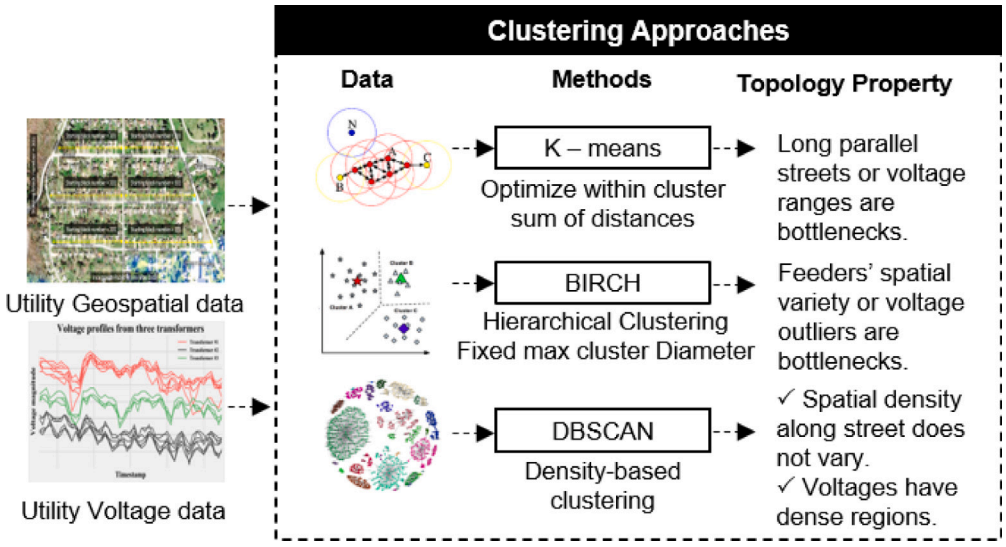


Fig. 1. Comparison of three important families of algorithms for clustering based on both GIS and AMI data.

the addresses using Bing Maps API to get the GPS coordinates for meters.

For temporal data from our partner electric utility, the raw data consists of a separate data file for each customer based on the utility collection mechanism. To process such temporal data, we need to merge the data in a tabular form, e.g., rows representing smart meters and columns representing timesteps. To combine the data files, we select the timesteps that are common in all the files. In addition, we removed all values that do not lie within the ±1% of the base values, as such measurements are more likely to be erroneous than the values in the normal range.

3.2. Metric evaluation for clustering algorithm design

For data clustering, there are numerous approaches. For these methods, three categories are popular in machine learning fields. One is to consider the group properties, e.g., calculate the sum of distances within each cluster. The second category is to bound the cluster with a limit, e.g., maximum diameter for clusters. The third category investigates the importance of cluster "density", e.g., the number of data points in a neighborhood of points. Fig. 1 provides the visual ideas of the three categories, and we analyze their typical algorithm's suitability for power data.

3.2.1. K-means for average distances

One idea for clustering is to consider the average distances for all the members in a group. For example, K-means was originally proposed in [25]. It is one of the common clustering methods that is applied in a variety of scenarios. For example, it has been applied to identify optimal placement of distributed generation sources in distribution systems [28], security assessment of power systems [29], and renewable power forecasting [30]. Also, K-means is critically evaluated in the literature. For example, [31] evaluated the performance of K-means, [32] evaluated the accuracy and running time of K-means, and [33] evaluated various initialization techniques for the K-means algorithm.

K-means creates k centroids $\overline{\mathbf{x}^j} = \frac{1}{n_j}\sum_{i\in cluster(j)}\mathbf{x}^i$, where n_j is the number of smart meters in cluster j. It aims at minimizing the square error loss

$$J = \sum_{j=1}^{k}\sum_{i\in cluster(j)}(\|\mathbf{x}^i - \overline{\mathbf{x}^j}\|)^2, \tag{1}$$

where $\|\mathbf{x}^i - \overline{\mathbf{x}^j}\|$ is the Euclidean distance between a point $\mathbf{x}^i$ and centroid $\overline{\mathbf{x}^j}$ iterated overall points in the j_{th} cluster, for all n clusters.

Drawback for our problem: While such a method can be used for clustering, determining the number of clusters beforehand would be a problem for distribution grids. Moreover, for the geographical space, the streets can be curved and may be of irregular shape due to the terrain. Even in the voltage space, the true clusters can have irregular shapes due to the feeder geometry, confusing K-means.

3.2.2. BIRCH for maximum cluster distance

Instead of looking at the grouping effects in K-means, one can also bound the extreme points, e.g., in BIRCH. The BIRCH method was proposed in [34]. It has been used in transformer health status monitoring in [35], improving the economy of power systems with high capacity thermal power [36], and scenario generation of wind power [37]. BIRCH has been evaluated for big data in [38], and its performance is compared with other clustering algorithms in [39].

BIRCH requires three parameters: the branching factor B_r, the threshold T, and the cluster count k. The cluster centers $\overline{\mathbf{x}^j} = \frac{1}{n_j}\sum_{i\in cluster(j)}\mathbf{x}^i$, where n_j is the number of smart meters in cluster j, and the cluster radii

$$R_j = \sqrt{\frac{1}{n_j}\sum_{i\in cluster(j)}(\mathbf{x}^i - \overline{\mathbf{x}^j})^2} \tag{2}$$

can then be computed for each cluster. Every point is assigned to the nearest-center sub-cluster.

Drawback for our problem: For distribution systems, the geographical radius can be different, e.g., long feeders and short feeders. Thus, it is hard to put a limit on the diameter for the geographical space and voltage space.

3.2.3. DBSCAN for local densities

In the two approaches above, the focuses are on either the group property or on the property of an extreme limit. Another idea is to focus on a subgroup of points and check how the trend is propagating, which is the third category. For example, DBSCAN (Density-based spatial clustering of applications with noise) forms clusters based on two parameters: 1) a neighborhood region specified by the radius ϵ and 2) the minimum number of data points $minPoints$ in the neighborhood. The algorithm counts the data points in the sphere of radius ϵ around a data point and includes it in the cluster if it exceeds $minPoints$.

The DBSCAN algorithm was originally proposed in [27]. Density-based clustering (DBSCAN) finds applications in various areas of power

systems. For example, DBSCAN is used for outlier detection in [40], for consumer behavior analysis in [41], and to detect grid voltage oscillatory modes with high amplitudes for corrective action.

Advantage for our problem: Since DBSCAN can identify clusters of any irregular shape, it is good for power system geographical data since the streets may have an irregular shape. Moreover, setting the parameter $minPoints = 1$ can avoid any smart meter being neglected as an outlier without affecting other clusters.

3.3. Proposed density-based method

3.3.1. Distance for geographical data and for voltage data

For GPS data, let $\mathbf{l}^1, \mathbf{l}^2 \in \mathbf{L}$ be two latitude–longitude pairs in radians. Their distance in km on Earth's surface is given by the Haversine formula

$$d_{\mathbf{L}}(\mathbf{l}^1, \mathbf{l}^2) = 2R_E \cdot \arcsin\left\{\sqrt{\sin^2\left(\frac{l_1^1 - l_1^2}{2}\right) + \cos\left(l_1^1\right)\cos\left(l_1^2\right)\sin^2\left(\frac{l_2^2 - l_2^1}{2}\right)}\right\}, \tag{3}$$

where $R_E = 6371$ km is the radius of Earth. For the distance in the voltage domain, we use mutual information to quantify the distance. Specifically, the voltage–distance between two points $\mathbf{v}^1, \mathbf{v}^2 \in \mathbf{V}$ is defined as $d_{\mathbf{V}}(\mathbf{v}^1, \mathbf{v}^2) = \frac{1}{I(\mathbf{v}^1, \mathbf{v}^2)}$ where $I(\mathbf{v}^1, \mathbf{v}^2)$ is the mutual information between $\mathbf{v}^1$ and $\mathbf{v}^2$. The key idea of mutual information-based topology analysis in the past is based on using voltage correlation in a probabilistic way [42]. A distribution system typically has a radial structure. Therefore, we can represent the voltage data in a graphical model via the joint probability density

$$P_{\mathbf{V}}(v^2, v^3, \ldots, v^N) = P_{\mathbf{V}}(v^2)P_{\mathbf{V}}(v^3|v^2)\cdots P_{\mathbf{V}}(v^N|v^2, \ldots, v^{N-1}), \tag{4}$$

where we assign the swing bus as bus 1 with a deterministic value, which is eliminated from the measurements.

Based on such a chain rule, mutual information can be used for measuring voltage similarity, e.g., in the discrete-time scenario, mutual information is defined as

$$I(\mathbf{v}^1, \mathbf{v}^2) = \sum_{i=1}^{T}\sum_{j=1}^{T} p_{(\mathbf{v}^1,\mathbf{v}^2)}(v_i^1, v_j^2)\ln\left(\frac{p_{(\mathbf{v}^1,\mathbf{v}^2)}(v_i^1, v_j^2)}{p_{\mathbf{v}^1}(v_i^1)p_{\mathbf{v}^2}(v_j^2)}\right). \tag{5}$$

Essentially, it is a weighted sum measuring the averaged similarity between the joint probability density $p_{(\mathbf{v}^1,\mathbf{v}^2)}(v_i^1, v_j^2)$ and the products of the individual probability densities, $p_{\mathbf{v}^1}(v_i^1)\cdot p_{\mathbf{v}^2}(v_j^2)$. For example, if v_i^1 and v_j^2 are independent random variables, $p_{(\mathbf{v}^1,\mathbf{v}^2)}(v_i^1, v_j^2) = p_{\mathbf{v}^1}(v_i^1)\cdot p_{\mathbf{v}^2}(v_j^2)$, making $\ln\left(\frac{p_{(\mathbf{v}^1,\mathbf{v}^2)}(v_i^1,v_j^2)}{p_{\mathbf{v}^1}(v_i^1)p_{\mathbf{v}^2}(v_j^2)}\right) = 0$ in Eq. (5), showing no connection between buses i and j. On the other hand, neighboring smart meters sharing a common transformer have similar voltage profiles resulting in high mutual information.

Based on the distances in the voltage and geographical domains, the combined distance of two datapoints $\mathbf{x}^1, \mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is given as $d_{[\mathbf{L},\mathbf{V}]}(\mathbf{x}^1, \mathbf{x}^2) = d_{\mathbf{L}}(\mathbf{l}^1, \mathbf{l}^2) + d_{\mathbf{V}}(\mathbf{v}^1, \mathbf{v}^2)$.

3.3.2. Evaluation of the density in the combined space of geographical and voltage data

To define a notion of density in $(n+1)$-dimensional space $[\mathbf{L}, \mathbf{V}]$, we first consider the two-dimensional space L. For argument's sake, consider two points $\mathbf{l}^1, \mathbf{l}^2 \in \mathbf{L}$. $\mathbf{l}^1 = (l_1^1, l_2^1)$, $\mathbf{l}^2 = (l_1^2, l_2^2)$ in a 2-D space. The Euclidean distance for these two points is $d(\mathbf{l}^1, \mathbf{l}^2) = [(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{0.5}$. If we fix the distances to be less than ϵ, then we obtain the following: $d(\mathbf{l}^1, \mathbf{l}^2) = [(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{0.5} < \epsilon$. Squaring both sides yields: $(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2 < \epsilon^2$. The equation looks similar to the equation of a circle with radius ϵ and center is at the point (l_1^2, l_2^2). Thus, the algorithm counts the data points in the sphere of radius ϵ around a data point and includes it as a core point in the cluster if it exceeds $minPoints$. However, using Euclidean distance is wrong due to Earth's spherical shape, and therefore, we use Haversine distance that gives the distance on the surface of Earth in km.

Definition 1 (*ϵ-neighborhood of a Point*)**.** The ϵ-neighborhood of a datapoint $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$, denoted by $N_r(\mathbf{x}^1)$, is defined by

$$N_r(\mathbf{x}^1) = \left\{\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}] : d_{[\mathbf{L},\mathbf{V}]}(\mathbf{x}^1, \mathbf{x}^2) < \epsilon\right\}. \tag{6}$$

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*. The points at the boundary of a cluster may not qualify to be a core point. For such points, we cluster them with a core point if they are in ϵ-neighborhood of a core point.

Definition 2 (*Directly Density-reachable*)**.** A point $\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is directly density-reachable from a point $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$ with respect to (w.r.t.) ϵ and $minPoints$, if (1) $\mathbf{x}^2 \in N_r(\mathbf{x}^1)$, and (2) $N_r(\mathbf{x}^1) \geq minPoints$ ($\mathbf{x}^1$ is a *core point*).

Directly density-reachability is not transitive. To ease algorithmic development, we need a transitive property.

Definition 3 (*Density-reachable*)**.** A point $\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is *density-reachable* from a point $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$ w.r.t. ϵ and $minPoints$, if there is a sequence of points $\mathbf{y}^1, \ldots, \mathbf{y}^m \in [\mathbf{L}, \mathbf{V}], \mathbf{y}^1 = \mathbf{x}^2, \mathbf{y}^m = \mathbf{x}^1$, so that $\mathbf{y}^{i+1}$ is directly density reachable from $\mathbf{y}^i$.

Definition 4 (*Density-connected*)**.** A point $\mathbf{x}^2$ is *density-connected* to a point $\mathbf{x}^1$ w.r.t. ϵ and $minPoints$ if there is a point $\mathbf{x}^3$ such that $\mathbf{x}^2$ and $\mathbf{x}^1$ are density-reachable from $\mathbf{x}^3$.

According to DBSCAN, two points are in the same cluster if and only if they are density connected. Density connectedness is a reflexive, symmetric, and transitive property. Therefore, it is guaranteed to form equivalence classes that are the clusters.

3.3.3. Density-based algorithm

We start with some point, $\mathbf{x}^1$, and check if it is a core point by the condition $N_r(\mathbf{x}^1) \geq minPoints$. For example, we can set $minPoints = 1$ to ensure that no smart meter is neglected as an outlier in the rural sparse distribution grid. Essentially, the distance between $\mathbf{x}^1$ and $\mathbf{x}^2$ is not the usual Euclidean distance but the specific Haversine distance. If $\mathbf{x}^1$ is a core point, we keep it as a starting point for the cluster. If $\mathbf{x}^1$ is not a core point, we put it in the outliers list and randomly select another point and repeat the procedure until we find a core point. In such a case, all of $N_r(\mathbf{x}^1)$ are in the same cluster as $\mathbf{x}^1$. Next, we individually check each point in $N_r(\mathbf{x}^1)$ for core point. All newly discovered core points are inserted in a queue. Next, we repeat the same procedure for each core point in the queue, thereby adding new points to the cluster and the core points queue until the core points queue is empty, making cluster one complete. Subsequently, we randomly start searching the remaining points for a new core point for the second cluster and repeat such a process. Algorithm 1 is different from the original DBSCAN [27] as it considers only the core points. Algorithm 2 is an improved version that is robust against adversarial noise [43]. Algorithm 2 calls Algorithm 1 in step 2.

Algorithm 1: Core_DBSCAN

Input: X, ϵ, $minpts = 1$

1 . $H := \{\mathbf{x} \in X : |B(\mathbf{x}, \epsilon) \cap X)| \geq minpts\}$.

2 $G :=$ undirected graph with vertices H. An edge between $\mathbf{x}, \mathbf{x}' \in H$ exists if $|\mathbf{x} - \mathbf{x}'| \leq \epsilon$.

3 **return** connected components of G.

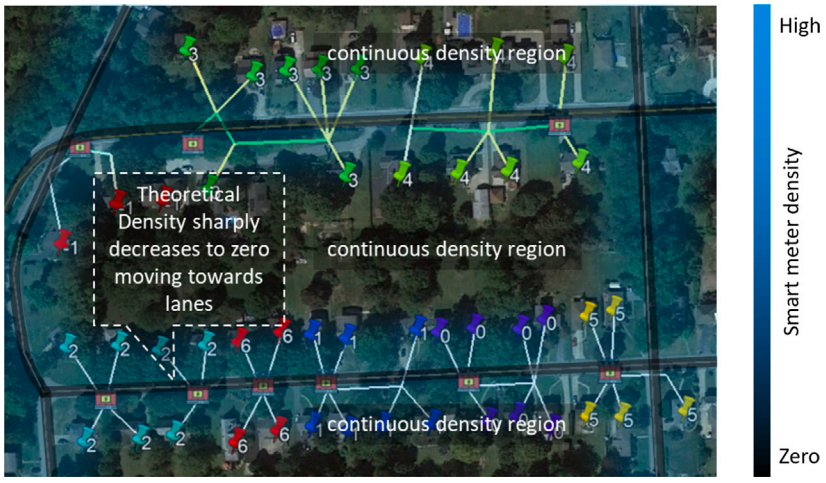


Fig. 2. Street lanes cannot have any smart meters, which causes a discontinuity at the street boundary, but a removal can remove such a discontinuity problem.

Algorithm 2: Robust DBSCAN

Input: X, ϵ, $\widetilde{\epsilon}$, $minpts = 1$

$H := \{\mathbf{x} \in X : |B(\mathbf{x}, \epsilon) \cap X)| \geq minpts\}$.

$D :=$ Core_DBSCAN(X, $\widetilde{\epsilon}$, $minpts$).

$C := \{C \cap H : C \in D\}$.

return C.

4. Guarantee of the density-based algorithm

In this section, we provide an explainable and intuitive theoretical guarantee for the Robust DBSCAN in Algorithm 2 and show that the algorithm is robust against the addition of new data. In particular, we show that adding l new utility customers with smart meter voltage and location data to the original data does not change the original clusters, and the cluster assignments to the original points remain unchanged, i.e., the original points that were clustered together (separate) remain together (separate) after adding new points.

4.1. The first assumption on differential density function

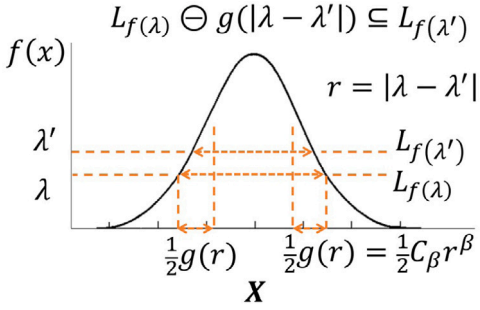
Assumption 1. The density function $f(\mathbf{x})$ should be differentiable.

The assumption is regarding the theoretical density $f(\mathbf{x})$ for a point $\mathbf{x}$ in the combined geographic-voltage space and not the measured density $N_r(\mathbf{x})$. Therefore, the assumption remains valid if the measured density is not differentiable. Even for spatial analysis with smart meter geographic density $f_g(\mathbf{x})$ equaling zero on the street lanes, as shown in Fig. 2, we can remove the street lanes from the domain of the density function to avoid a step change discontinuity. In order to have a mathematical analysis of the density, we need to define *superlevel-set $L_f(\lambda)$ of the density function f corresponding to a given threshold (level) λ* as a set of all points in the dataset $[\mathbf{L}, \mathbf{V}]$ with a density equal to or greater than the threshold λ. Moreover, if Assumption 1 holds, the superlevel-sets consist of closed intervals rather than discrete points.

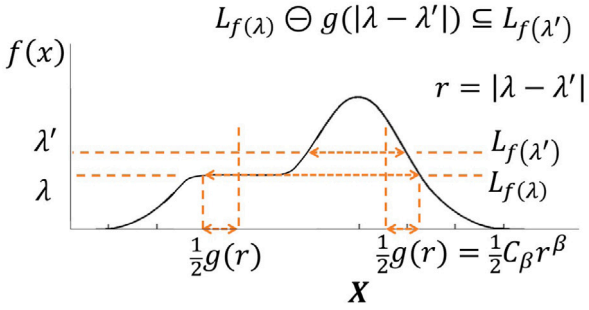
4.2. The second assumption on curvature

Usually, the shape of a density function has one or more overlapping bell curves or some flat regions. Therefore, if we have two levels, λ and λ', such that $0 < \lambda \leq \lambda' < \|f\|_\infty$, where $\|f\|_\infty$ represents the peak density, then the superlevel-set for level λ' is a subset of the superlevel-set for level λ. Mathematically, $L_f(\lambda') \subseteq L_f(\lambda)$.

Given a continuous set A, if we "trim" set A from all sides of the boundary by a depth δ, the remaining set is called the *δ-interior of A*.



(a) A probability density function that satisfies Assumption 2.



(b) A probability density function with a strictly positive flat region does not satisfy Assumption 2.

Fig. 3. Examples of probability density functions based on Assumption 2.

For example, in Fig. 3(a), we can "trim" $L_f(\lambda)$ from its boundary by a depth g to make it a subset of $L_f(\lambda')$. Mathematically, we denote "trimming" a depth g from all boundaries of a superlevel-set $L_f(\lambda)$ as $L_f(\lambda) \ominus g$ [43]. Such a concept is the basis for Assumption 2.

To provide a guarantee for the robustness of density-based clustering, we need the density function to decay around the cluster boundaries so that the clusters are salient enough to be detected. In particular, we need no strictly positive flat regions in the smart meter density function. Strictly positive flat regions in the density function can be avoided in the following way. For Fig. 3(a), assume g is an increasing homogeneous function of $(\lambda' - \lambda)$ and assume for all $0 < \lambda \leq \lambda' < \|f\|_\infty$, where $\|f\|_\infty$ represents the peak density, we have $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$. This is because there is no strictly positive flat region in Fig. 3(a). However, for Fig. 3(b), if we set λ and λ' just below

and above the flat region, we cannot obtain $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$ due to the strictly positive flat region. Assumption 2 below gives a formal description of this concept. Flat regions in density function with a zero value satisfy Assumption 2, as the superlevel-set $L_f(\lambda = 0)$ is not included in the assumption.

Assumption 2 (*Curvature*)**.** There exists $C_\beta > 0$ and $\beta > 0$ such that the following holds. For any $0 < \lambda < \lambda' < \|f\|_\infty$, we have $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$ where $g(r) = C_\beta \cdot r^\beta$.

In the voltage domain, a density function satisfying Assumption 2 means that the voltage distances (inverse of the mutual information) gradually increase as we move from the center of clusters (houses supplied by the same transformers) to external areas. The function $g(r) = C_\beta \cdot r^\beta$ mimics the exponential structure. For example, the family of exponential probability densities can estimate many real-world phenomena. Assumption 2 ensures that the density function is not strictly positive and flat simultaneously. Moreover, Assumption 2 forces sufficient density decay around the superlevel-sets so that the superlevel-sets are salient and will be detected [43]. We can introduce a slightly different density estimator concept than $N_r(\mathbf{x})$, i.e., to keep the number of points k fixed and adjusting the radius $r_k(\mathbf{x})$ to enclose k nearest neighbor points with the sphere called the k-NN density estimator. A lot of literature is based on this approach, formally defined as $f_k(\mathbf{x}) := \frac{k}{n \cdot v_D \cdot r_k(\mathbf{x})^D}$, where v_D is the volume of a unit ball in $d_{[\mathbf{L},\mathbf{V}]}$, $r_k(\mathbf{x})$ is the adjusted radius of the sphere to enclose k points. $v_D \cdot r_k(\mathbf{x})^D$ is the volume of the sphere with radius $r_k(\mathbf{x})$, and $\frac{k}{v_D \cdot r_k(\mathbf{x})^D}$ is the number of points per unit volume. In order to remove the effect of the total number of points n, we divide it by n. Once we have the required assumptions and definitions, now we can go ahead with the proof.

4.3. Robustness guarantees against meter number

We now show robustness guarantees on the core points returned by Algorithm 2. In particular, we show that adding l new utility customers with smart meter voltage and location data to the original data does not change the original clusters. The cluster assignments to the original points remain unchanged i.e., the original points that were clustered together (separate), remain together (separate) after adding new points. That is, when running Algorithm 2 on $[\mathbf{L}, \mathbf{V}]$ vs. running it on $[\mathbf{L}', \mathbf{V}']$ with l additional samples, any new core points that appear will be near the original core points.

The k-NN density estimation error can be given by a probabilistic bound between the true density $f(\mathbf{x})$ and the k-NN density estimation $f_k(\mathbf{x})$. Such a bound can be used to identify the upper bound of the theoretical density given the k-NN density estimation via density-based clustering. The upper bound of the true density can be used to provide a guarantee for core points. For measuring $f_k(\mathbf{x})$, if k is very small, it can lead to estimation errors due to less samples within the sphere, reducing the estimation accuracy. Therefore, to provide a confidence level $(1 - \delta)$ for the bound, one needs to have a lower bound on k. The lower bound on k is directly related to the sample size n. Moreover, k is directly related to the confidence level $(1 - \delta)$. Lemma 1 directly follows from Lemma 3 and 4 of [43,44].

Lemma 1 (*k-NN Density Estimation Accuracy*)**.** *Let $0 < \delta < 1$. Suppose that f satisfies Assumption 1. Then the following holds for some constants C and C_l depending on f. Suppose k satisfies $k \geq C_l \cdot \log(\frac{1}{\delta^2}) \cdot \log n$. Then with a probability of at least $1 - \delta$, the following holds:*

$$\sup_{\mathbf{x} \in [\mathbf{L},\mathbf{V}]} |f(\mathbf{x}) - f_k(\mathbf{x})| \leq \left(\frac{\log(\delta^{-1}\sqrt{\log n})}{\sqrt{k}} + \left(\frac{k}{n}\right)^{\frac{\alpha}{D}} \right). \tag{7}$$

Lemma 1 provides the limit to the error in the k-NN estimator accuracy $f_k(\mathbf{x})$. Indeed, the range of error is directly related to the confidence level $(1-\delta)$. Also, a greater sample size n can lead to greater error if k is small since the number of points within the sphere will be even smaller as compared to the total sample size n. Moreover, a higher degree of continuity α of the density function will result in a lower error. From Assumption 1, we have that the density function is continuous. Moreover, from Assumption 2, we have that the density function has a curvature and is never flat. Furthermore, using Lemma 1, we have that the true density will not be much different than measured by DBSCAN. Therefore, the new l points will lie close to the original clusters. In fact, using the above three arguments, we can calculate the probabilistic maximum extension $\widetilde{r}$ from the original DBSCAN clusters. Therefore, the new clusters C' will be bounded by the original clusters extended by the distance $\widetilde{r}$ with a confidence of $1-\delta$. The lower bound on k remains the same as Lemma 1. However, the total number of points becomes $(n + l)$.

Lemma 2. *Suppose that Assumptions 1 and 2 hold. There exists constants C_l and C depending on f such that the following holds. Let $0 < \delta < 1$ and k satisfy $k \geq C_l \cdot \log(\frac{1}{\delta^2}) \cdot \log(n + l)$, and $\widetilde{\epsilon} > \epsilon > 0$. Let C and C' be the core points returned by Algorithm 2 when run on $[\mathbf{L}, \mathbf{V}]$ and $[\mathbf{L}', \mathbf{V}']$, respectively. With probability at least $1 - \delta$, the following holds: $C' \subset C \oplus \widetilde{r}$, where $\oplus$ denotes a tube around a set (i.e. $A \oplus r := \{\mathbf{x} \in [\mathbf{L}, \mathbf{V}] : \inf_{a \in A} |\mathbf{x} - a| \leq r, \}$) and $\widetilde{r} < \infty$.*

The proof of Lemma 2 follows Assumptions 1, and 2, and Lemma 1 [43]. The result $C' \subset C \oplus \widetilde{r}$ suggests that the new points lie within the tube of thickness $\widetilde{r}$ around the original clusters. Therefore, if the edges of the original clusters are at a distance $2\widetilde{\epsilon} + 2\widetilde{r}$, there will not be any original clusters merging to form one cluster. Moreover, if $\widetilde{r} < \widetilde{\epsilon}$, then the new points will not form separate clusters.

Theorem 1. *Suppose that conditions of Lemma 2 hold. Let C, C' be the output of Algorithm 2 on $[\mathbf{L}, \mathbf{V}]$ and $[\mathbf{L}', \mathbf{V}']$, respectively, and define the minimum inter-cluster distance of the returned clusters*

$$R := \min_{C_1, C_2 \in C, C_1 \neq C_2} \min_{\mathbf{x}^1 \in C_1, \mathbf{x}^2 \in C_2} d_{[\mathbf{L},\mathbf{V}]}(\mathbf{x}^1, \mathbf{x}^2). \tag{8}$$

If additionally, the following holds: $\widetilde{r} \leq \widetilde{\epsilon} \leq \frac{1}{2}R - \widetilde{r}$, then $|C| = |C'|$ (i.e. the number of clusters does not change) and there exists a one-to-one mapping of the clusters $\sigma : C \to C'$ such that $C \subset \sigma(C)$ for all $C \in C$ (i.e., original clusters are preserved).

Proof. Note that all the points appearing in a cluster of C will also appear in some cluster of C'. By Lemma 1, we have that any newly appearing points in C' will be at a distance of at most $\widetilde{r}$ from a point appearing originally in C, mathematically $C' \subset C \oplus \widetilde{r}$. From the assumption $\widetilde{\epsilon} \geq \widetilde{r}$, we have that the radius hyperparameter for DBSCAN is lesser than $\widetilde{r}$, then such new points will become reconnected to the same cluster in C since they will be present in the sphere of radius $\widetilde{r}$. Finally, from the assumption $\widetilde{\epsilon} \leq \frac{1}{2}R - \widetilde{r}$, we have that the original clusters are separate by more than $2\widetilde{\epsilon} + 2\widetilde{r}$, which means that no two distinct clusters in C will become merged in C'. □

5. Deployment

Fig. 4 shows a deployment of the proposed algorithm in our utility partner's territory. It is obtained by directly running the algorithm on the sets of data without any human intervention. For example, in the middle one, the green lines show the connections from poles to poles, the yellow lines show the connections from the poles to smart meters, and the red one is the primary feeder topology. This visualization shows that such an algorithm is suitable for large-scale topology recovery.

Fig. 4. Demo of deployed topology clustering algorithm.

6. Numerical validation

6.1. Data description

The simulations are implemented on the IEEE PES distribution networks for IEEE benchmark systems, such as 123-bus systems. We also implement our algorithm on a utility grid. For benchmark systems, the feeder bus is selected as the slack bus. To simulate the power system behavior in a more realistic pattern, the load profiles from Pacific Gas and Electric Company (PG&E) and "ADRESConcept" Project of Vienna University of Technology [45] are adopted as the real power profile in the subsequent simulation. PG&E load profile contains hourly real power consumption of 123,000 residential loads in northern California, USA. "ADRES-Concept" Project load profile contains real and reactive powers profile of 30 houses in Upper-Austria. The data were sampled hourly over 14 days, so we generate voltage data using the historical consumption data with load flow analysis by the MATPOWER and OpenDSS.

For the utility grid, it is a mid-sized northeast system that includes approximately 600,000 customers, 7200 miles of overhead conductors, 250,000 poles, 108,000 transformers, 4500 miles of cable, 1000 sectionalizers, 400 capacitors, and 500 network protectors. A sample of 10,000 customers' AMI voltage data was used as well as the nearby transformers' GPS coordinates and the GPS coordinates of the poles. A summary of the voltage information is shown in Table 1.

6.2. Robust clustering

6.2.1. Validation on IEEE-123 test case system

As public secondary distribution is hard to find, we adjust IEEE 123-bus test system by adjusting its line parameters, e.g., R/X ratio, into the parameter range of the secondary grids. Afterwards, we randomly cut and separated the system into two systems, each with its own transformer, shown in Fig. 5, so that the validation process is free from any biases and for extensive testing. Since there is no test system available for the secondary distribution system, we use the IEEE 123-bus test system. Moreover, we change the line parameters of the IEEE 123-bus test system to mimic a secondary distribution feeder. For example, we change the X/R ratio of the lines to 0.2, which is a typical value for the X/R ratio of low voltage cables.

The system is disconnected at any bus to create two separate subsystems, e.g., split between bus 67 and 68 to provide an even split in Fig. 5. Afterwards, we run load flow analysis on 500 load scenarios for each split to generate a voltage dataset over a typical load cycle, similar to what the utility provided us. As the IEEE 123-bus test system does not provide any GPS coordinates of the nodes, we use the location coordinates using coordinates from the OpenDSS IEEE-123 bus model. With voltage data and "GPS coordinates", we run the proposed clustering algorithm.

To understand the performance of our clustering process, we compare two other clustering methods to our proposed clustering method. The input to BIRCH and Kmeans is the voltage time-series, while

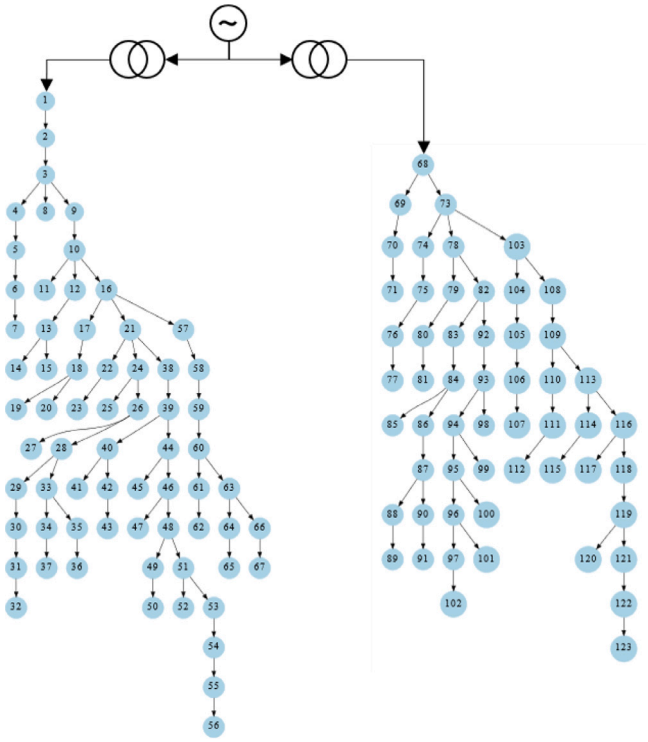


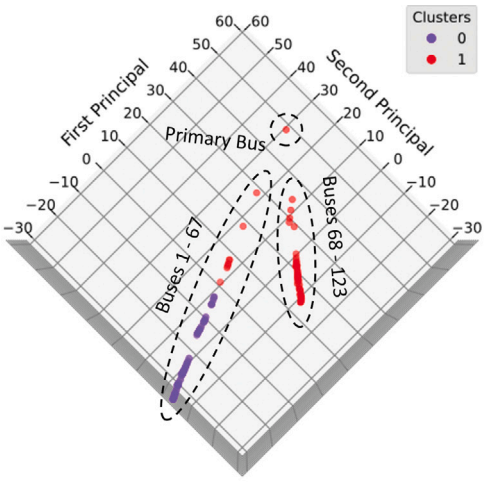
Fig. 5. Network Partition. The IEEE 123-bus system was used to understand the different dynamics of the three clustering algorithms for illustration purposes.

Table 1
Voltage data summary provided by utility partner.

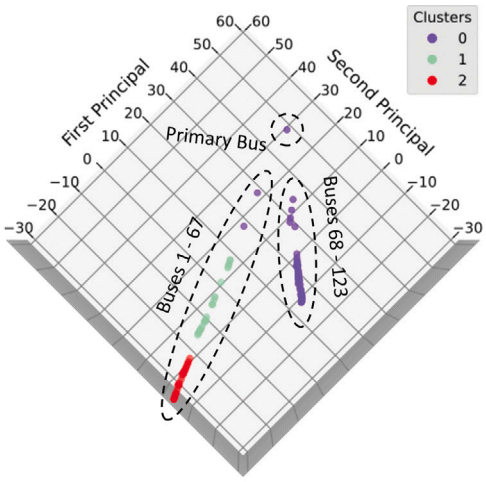
	Area 1	Area 2	Area 3
Total number	3442 × 8640	2578 × 8640	2357 × 8640
MAC addresses	3442	2578	2357
Starting time	2016/7/22	2016/7/22	2016/7/22
Ending time	2016/8/21	2016/8/21	2016/8/21
Units	Volt	Volt	Volt

DBSCAN uses mutual information of voltage time-series designed in this paper. This is due to the ability of DBSCAN to utilize the mutual information by considering mutual information as an inverse of the distance in feature space, while Kmeans and BIRCH cannot utilize mutual information. Moreover, since Kmeans also need the number of clusters while DBSCAN and BIRCH do not need it, we specify $k = 2$ for Kmeans as there are two transformers as prior information.

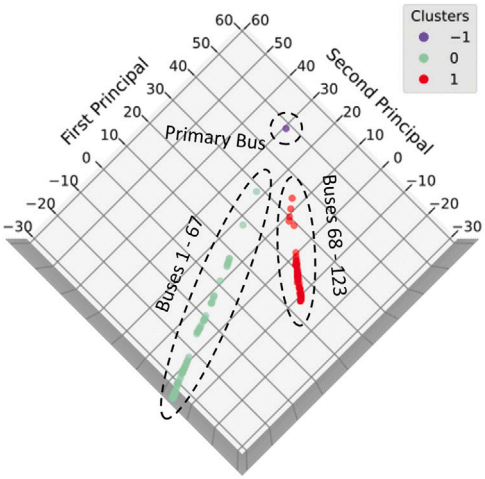
The result of the comparison is shown in Fig. 6. The Kmeans algorithm divides the network based on the minimum sum of within-cluster centroid distances. As can be observed from Fig. 6a, the sum of within-cluster centroid distances will be higher for the ground truth,



(a) Result of Kmeans clustering algorithm.



(b) Result of BIRCH clustering algorithm.



(c) Result of our proposed algorithm is the same as the ground truth of the modified IEEE-123 bus feeder, as shown in Fig. 5. The cluster −1 represents the outliers.

Fig. 6. Comparison of the three clustering algorithms using voltage and location information on IEEE-123 bus test feeder.

i.e., forming buses $1-67$ into a cluster. It is because the ground truth clusters have less gap between them near the primary bus. Moreover, considering the usual scenario of one way power flow, the ground truth clusters only gradually separate, moving towards the ends of the feeders, as shown in Fig. 6. To minimize the sum of within-cluster centroid distances, the Kmeans algorithm introduces an error, i.e., splitting the ground truth cluster of buses $1-67$. Therefore, the approach of minimizing the sum of within-cluster centroid distances (i.e., the Kmeans algorithm) is not suitable for splitting the utility data.

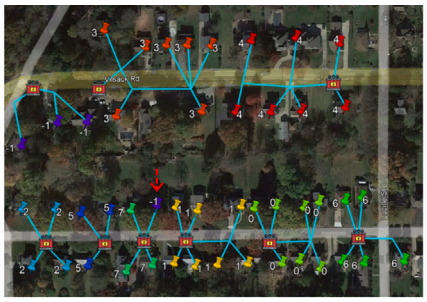
Similar logic can be used for the BIRCH algorithm since it imposes a strict diameter threshold on the clusters, as can be observed in Fig. 6b. Moreover, for BIRCH, the order of data presentation also matters, resulting in different clusters if it begins from the root or leaf nodes. In this example, cluster 0 grows until it reaches the threshold. The algorithm abruptly stops adding more points to cluster 0 and starts cluster 1.

By comparing Fig. 6c with the other two subfigures, we can see that only our density-based method is clustering consistently. Such an observation remains when we change the loads and topology, showing the power of integrated design of the machine learning method with the needs of power systems. It is because a density-based method starts with a high-density point (core-point) and progresses through high-density regions until it finds a break due to the two transformer impedances. It is how the ground truth clusters are formed in the voltage feature space, as can be observed in Fig. 6c. With hyper-parameter tuning, such a method can segment the utility data effectively.

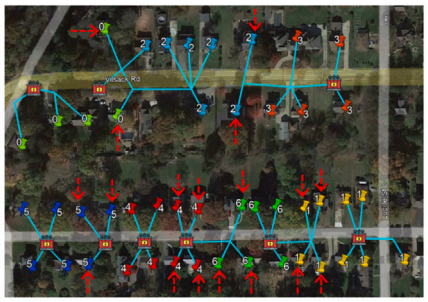
We also observe that more data availability increases the accuracy, where Table 2 presents a numerical comparison of accuracy versus the data availability, where we tune the hyperparameters of DBSCAN to maximize accuracy for each value of data availability. Moreover, we add the execution times, mean absolute error, and mean square error of the algorithm. With the exception of a few samples, i.e., 10, we see a straight relationship between the samples and the execution time. In the case of few samples, the density-based method considers the majority of data points as outliers. Outliers are rechecked once clusters are finalized to verify if they correspond to a cluster's border. As a result, the time necessary to execute 10 samples is longer than the time required to execute more samples.

Finally, we put our method to the test with noisy data. The supplied data is contaminated by 0.1% noise. For example, we multiply a standard normal distribution by 0.1% of the input voltage dataset's mean value. As indicated in Table 2, the results remain mainly unaffected for larger numbers of samples.

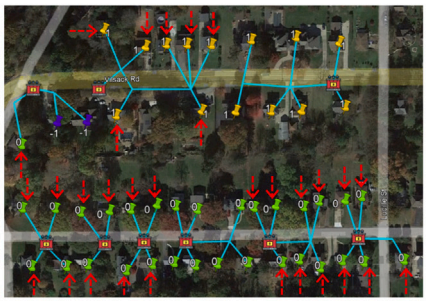
(a) Kmeans clustering using Voltage information only.



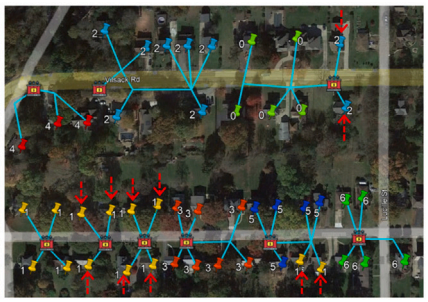
(b) DBSCAN clustering using Voltage information only.



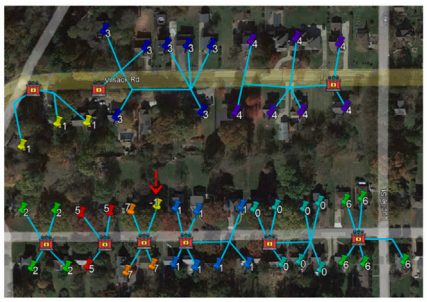
(c) Kmeans clustering using location information only.



(d) DBSCAN clustering using GIS information.



(e) Kmeans clustering using Voltage and location information.



(f) DBSCAN clustering using Voltage and location information.

Fig. 7. Comparison of the three clustering algorithms using voltage and location information on a sample in our partner utility.

Table 2
Voltage sample versus accuracy.

Samples	10	50	75	100	150	250	500
Execution time [s]	11.99	5.32	5.42	5.53	5.55	5.93	6.10
Accuracy	1%	36%	95%	100%	100%	100%	100%
Mean absolute error (MAE)	0.99	0.64	0.05	0	0	0	0
Mean square error (MSE)	0.99	0.64	0.05	0	0	0	0
Accuracy with 0.1% Noise data	0%	0%	54%	54%	54%	98%	96%

6.2.2. Validation on real utility system

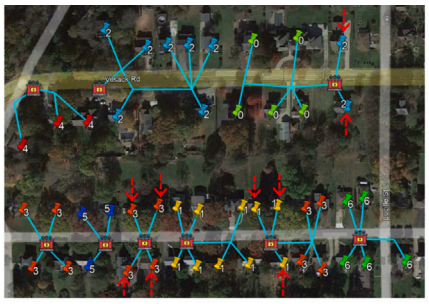
As our utility partner provides geographical location information, we also use real GPS data for validation of algorithmic results. To validate the importance of using the combined voltage-location dataset, we compare the results of two algorithms with (i) voltage dataset only, (ii) location dataset only, and (iii) combined voltage-location dataset, as shown in Fig. 7. Moreover, to validate the importance of using our proposed method, we compare our method to four other algorithms using the combined voltage-location dataset in Fig. 8. The scenarios also include comparisons with both recent and classical methods. However, we observe that our proposed method with results in Fig. 8(e) is the best with a consistent segmentation of transformer to meter connectivity among all combinations. Our algorithm is also the best according to the accuracy in Table 3, due to its capability to integrate the voltage information and ground distance in the best way.

6.3. Overall accuracy

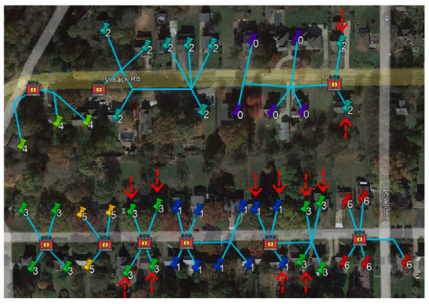
To evaluate the accuracy, we conduct our algorithm throughout the utility territory in Fig. 8. For methodology, we compare our algorithm with respect to a mutual information method with the Chow-Liu algorithm, the BIRCH method, and the k-means method. The results are displayed in Fig. 9, where the proposed method has an accuracy of near 95% over a large number of buses. The result is also quite robust if the bus number continues to grow. Moreover, we can see that for the large area, as shown in Fig. 9, location information improves the accuracy since voltages can be similar for smart meters over large areas due to similar neighborhood consumption profiles.

Remark 1. To deploy our method for large-scale validation, we assign addresses to poles to reverse-geocode each pole, which requires purchasing Google Maps API. With the parent transformer, poles, and the smart meters belonging to the same cluster, we can use a minimum spanning tree to connect them to obtain the secondary overhead

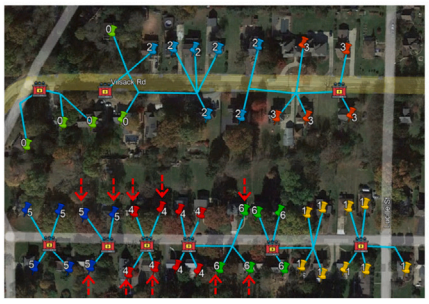
(a) Kmeans clustering using Voltage and location information.



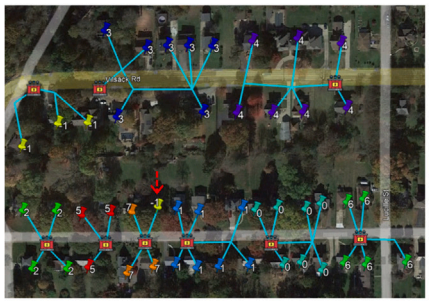
(b) BIRCH clustering using Voltage and location information.



(c) Hierarchical clustering [46] using Voltage and location information.



(d) InfleCS [47] clustering using Voltage and location information.



(e) DBSCAN clustering using Voltage and location information.

Fig. 8. Comparison of Kmeans, BIRCH, and our proposed density-based clustering algorithms using voltage and location information on a sample in our partner utility.

Table 3
A numerical comparison of methods on various datasets shown in Fig. 8 based on five reference metrics: accuracy, execution times, MSE, MAE, and AMI score.

Algorithm and data	Accuracy [%]	Execution Times [s]	MSE	MAE	AMI Score
Kmeans (Voltage Info.)	55.56%	0.26	0.44	0.44	0.768
BIRCH (Voltage Info.)	66.67%	0.01	0.33	0.33	0.751
Kmeans (GIS Info.)	33.33%	0.37	0.67	0.67	0.656
BIRCH (GIS/Voltage Info.)	66.67%	0.73	0.33	0.33	0.751
Kmeans (GIS/Voltage Info.)	55.56%	0.07	0.44	0.44	0.768
InfleCS [46] (GIS/Voltage Info.)	10.64%	28.6	0.89	0.89	−0.03
Hierarchical Clustering [47] (GIS/Voltage Info.)	80.85%	0.01	0.19	0.19	0.751
DBSCAN (GIS and Mutual Info of Voltage)	97.87%	1.91	0.02	0.02	0.966

MSE: mean square error; MAE: Mean absolute error; AMI: adjusted mutual information; InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures.

connections. Minimum spanning tree works by connecting the houses with the poles and transformers by minimizing the total length of wire. Such an algorithm is correct as (1) houses are usually supplied from their nearest poles without any measurements, and (2) the distribution system has a tree structure.

7. Conclusion

Electric utilities typically do not have an accurate distribution system topology readily available. With the advent of DERs, the electric utility faces challenges in the distribution grid. These challenges need greater visibility of their distribution system circuits through state

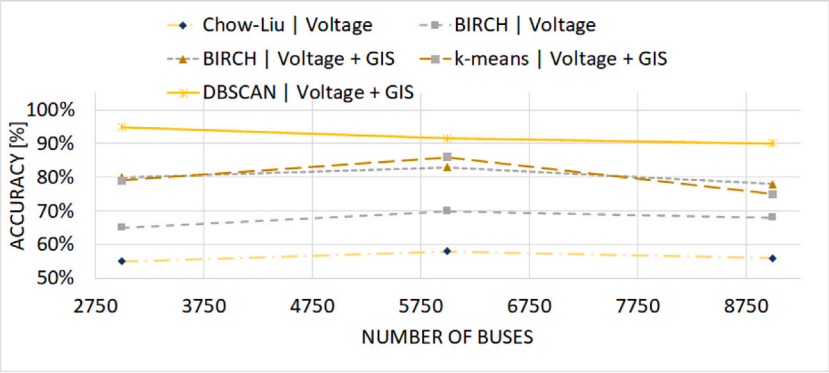


Fig. 9. Accuracy comparison for the whole utility areas.

estimation, where real-time topology recovery is the basis for modeling. Past technology is based on either outdated maps or use temporal information only and ignore the geographic information. However, temporal information is not enough for associating transformers with smart meters. This paper resolves this challenge by accurately clustering the topology. Specifically, we propose a density-based clustering method that leverages both voltage and geographical space data. And we show how to use GPS coordinates with voltage information to refine the connectivity within one transformer. Finally, we not only show how to improve our method but also provide an explainable theoretical bound. The proposed method is validated on the IEEE-123 bus system and the real system from our partner utility.

The proposed method does not require PMUs, a large amount of sensors as in transmission system, or voltage measurements at poles or transformers. Moreover, it can even work on areas having streets with irregular shapes. However, the proposed method requires both voltage and geographic data. Moreover, the guarantee for the proposed method exists under the conditions mentioned in the paper.

CRediT authorship contribution statement

Elizabeth Cook: Data curation, Software, Validation, Writing – original draft, Visualization. **Muhammad Bilal Saleem:** Proving the Guarantee, Writing – editing, Validation. **Yang Weng:** Supervision, Writing – editing, Conceptualization, Investigation, Reviewing. **Stephen Abate:** Reviewing and editing. **Katrina Kelly-Pitou:** Reviewing and editing. **Brandon Grainger:** Supervision, Reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Department of Energy under grants DE-AR00001858-1631 and DE-EE0009355, the National Science Foundation (NSF) under the grants ECCS-1810537 and ECCS-2048288.

References

[1] Kundur P, Balu NJ, Lauby MG. Power system stability and control. McGraw-hill New York; 1994.
[2] Guo Z, Zhou Z, Zhou Y. Impacts of integrating topology reconfiguration and vehicle-to-grid technologies on distribution system operation. IEEE Trans Sustain Energy 2019.
[3] Wood AJ, Wollenberg BF, Sheblé GB. Power generation, operation, and control. John Wiley & Sons; 2013.
[4] Lugtu R, Hackett D, Liu K, Might D. Power system state estimation: Detection of topological errors. IEEE Trans Power Appar Syst 1980.
[5] Weng Y, Negi R, Ilić MD. A search method for obtaining initial guesses for smart grid state estimation. In: International conference on smart grid communications. 2012.
[6] Korres GN, Manousakis NM. A state estimation algorithm for monitoring topology changes in distribution systems. In: Power and energy society general meeting. 2012.
[7] Huang J, Gupta V, Huang Y-F. Electric grid state estimators for distribution systems with microgrids. In: Annual conference on information sciences and systems. 2012.
[8] Zhang G, Lee S, Carroll R, Zuo J, Beard L, Liu Y. Wide area power system visualization using real-time synchrophasor measurements. In: IEEE PES general meeting. 2010.
[9] Deka D, Chertkov M, Backhaus S. Topology estimation using graphical models in multi-phase power distribution grids. IEEE Trans Power Syst 2019.
[10] Zamzam AS, Fu X, Sidiropoulos ND. Data-driven learning-based optimization for distribution system state estimation. IEEE Trans Power Syst 2019.
[11] Cavraro G, Arghandeh R, Barchi G, von Meier A. Distribution network topology detection with time-series measurements. In: IEEE international innovative smart grid technologies conference. 2015.
[12] Von Meier A, Culler D, McEachern A, Arghandeh R. Micro-synchrophasors for distribution systems. In: IEEE international innovative smart grid technologies conference. 2014.
[13] Weng Y, Rajagopal R. Probabilistic baseline estimation via gaussian process. In: Power & energy society general meeting. 2015.
[14] Weng Y, Negi R, Ilić MD. Historical data-driven state estimation for electric power systems. In: Inter. conf. on smart grid comm.. 2013.
[15] Liao Y, Weng Y, Rajagopal R. Urban distribution grid topology reconstruction via Lasso. In: Power & energy soc. gen. meeting. 2016.
[16] Fajardo OF, Vargas A. Reconfiguration of mv distribution networks with multi-cost and multipoint alternative supply, part ii: Reconfiguration plan. IEEE Trans Power Syst 2008.
[17] Baalbergen F, Gibescu M, van der Sluis L. Modern state estimation methods in power systems. In: IEEE/PES power systems conference and exposition. 2009.
[18] Yu J, Weng Y, Tan C-W, Rajagopal R. Probabilistic estimation of the potentials of intervention-based demand side energy management. In: International conference on smart grid communications. 2015.
[19] Baran ME, Jung J, McDermott TE. Topology error identification using branch current state estimation for distribution systems. In: Transmission & distribution conf. & exposition. 2009.
[20] Bolognani S, Bof N, Michelotti D, Muraro R, Schenato L. Identification of power distribution network topology via voltage correlation analysis. In: Conference on decision and control. 2013.
[21] Weng Y, Liao Y, Rajagopal R. Distributed energy resources topology identification via graphical modeling. IEEE Trans Power Syst 2017;32(4):2682–94.
[22] Yu J, Weng Y, Rajagopal R. PaToPaEM: A Data-driven parameter and topology joint estimation framework for time-varying system in distribution grids. IEEE Trans Power Syst 2019.
[23] Luan W, Peng J, Maras M, Lo J, Harapnuk B. Smart meter data analytics for distribution network connectivity verification. IEEE Trans Smart Grid 2015;6(4):1964–71.
[24] Blakely L, Reno MJ. Identifying errors in service transformer connections. IEEE power & energy society general meeting 2020;1–5.
[25] Macqueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of Berkeley symposium on mathematics statistics and probability, 1967.
[26] Zhang T, Ramakrishnan R, Livny M. BIRCH: AN efficient data clustering method for very large databases. Spec Interest Group Manage Data Rec 1996.
[27] Ester M, Kriegel H-P, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Knowledge discovery and data mining. 1996.

[28] Scarlatache F, Grigoraş G, Chicco G, Cârţină G. Using k-means clustering method in determination of the optimal placement of distributed generation sources in electrical distribution systems. In: 2012 13th international conference on optimization of electrical and electronic equipment (OPTIM). IEEE; 2012, p. 953–8.
[29] Kalyani S, Swarup KS. Particle swarm optimization based K-means clustering approach for security assessment in power systems. Expert Syst Appl 2011;38(9):10839–46.
[30] Sun Z, Zhao S, Zhang J. Short-term wind power forecasting on multiple scales using VMD decomposition, K-means clustering and LSTM principal computing. IEEE Access 2019;7:166917–29.
[31] Ahmed M, Seraj R, Islam SMS. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics 2020;9(8):1295.
[32] Singh N, Singh D. Performance evaluation of k-means and heirarichal clustering in terms of accuracy and running time. IJCSIT) Int J Comput Sci Inf Technol 2012;3(3):4119–21.
[33] Steinley D, Brusco MJ. Initializing K-means batch clustering: A critical evaluation of several techniques. J Classification 2007;24(1):99–121.
[34] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. ACM Sigmod Rec 1996;25(2):103–14.
[35] Chu Z, Wang W, Li B, Jin W, Liu S, Zhang B, Lin Z. An operation health status monitoring algorithm of special transformers based on BIRCH and Gaussian cloud methods. Energy Rep 2021;7:253–60.
[36] Du H, Li Y. An improved BIRCH clustering algorithm and application in thermal power. In: 2010 international conference on web information systems and mining, Vol. 1. IEEE; 2010, p. 53–6.
[37] Li Q, Tang X, Chen C, Liu X, Liu S, Shi X, Liu H, Li Z, Lin Z, Yang L, et al. BIRCH Algorithm and wasserstein distance metric based method for generating typical scenarios of wind power outputs. In: 2019 IEEE innovative smart grid technologies-Asia (ISGT Asia). IEEE; 2019, p. 3640–4.
[38] Akshatha S, Kumar N. Evaluation of BIRCH clustering algorithm for big data 4. 2019.
[39] Nayyar A, Puri V. Comprehensive analysis & performance comparison of clustering algorithms for big data. Rev Comput Eng Res 2017;4(2):54–80.
[40] Zhang P, Wang Y, Liang L, Li X, Duan Q. Short-term wind power prediction using GA-BP neural network based on DBSCAN algorithm outlier identification. Processes 2020;8(2):157.
[41] Zhang L, Deng S, Li S. Analysis of power consumer behavior based on the complementation of K-means and DBSCAN. In: 2017 IEEE conference on energy internet and energy system integration (EI2). IEEE; 2017, p. 1–5.
[42] Liao Y, Weng Y, Wu M, Rajagopal R. Distribution grid topology reconstruction: An information theoretic approach. In: North American power symposium. 2015.
[43] Jiang H, Jang J, Nachum O. Robustness guarantees for density clustering. In: Inter. conf. on artificial intelligence and stats.. 2019.
[44] Dasgupta S, Kpotufe S. Optimal rates for k-nn density and mode estimation. In: Advances in neural infor. processing sys.. 2014.
[45] Institute of Energy Systems and Electrical Drives. Adres-dataset. 2016, http://www.ea.tuwien.ac.at/projects/adresconcept/EN/.
[46] Westerlund AM, Delemotte L. InfleCS: Clustering free energy landscapes with Gaussian mixtures. J Chem Theory Comput 2019;15(12):6752–9.
[47] Chami I, Gu A, Chatziafratis V, Ré C. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. Adv Neural Inf Process Syst 2020;33:15065–76.