Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data

Wenyu Wang, Nanpeng Yu, Brandon Foggo
Department of Electrical and
Computer Engineering
University of California, Riverside
Riverside, California 92521
Email: wwang032@ucr.edu, nyu@ece.ucr.edu, bfogg@ucr.edu

Joshua Davis
Advanced Technology Labs
Southern California Edison
Westminster, California 92683
Email: joshua.davis@sce.com

***Abstract*—Accurate network and phase connectivity models are crucial to distribution system analytics, operations and planning. Although network connectivity information is mostly reliable, phase connectivity data is typically missing or erroneous. In this paper, an innovative phase identification algorithm is developed by clustering of voltage time series gathered from smart meters. The feature-based clustering approach is adopted where principal component analysis is first carried out to extract feature vectors from the raw time series. A constrained k-means clustering algorithm is then executed to separate customers/smart meters into various phase connectivity groups. The algorithm is applied on a real distribution feeder in Southern California Edison's service territory. The accuracy of the proposed algorithm is over 90%.**

***Keywords*—data mining, k-means clustering, phase identification, principal component analysis, smart meter.**

I. Introduction

Driven by stricter environmental regulations, technological advances, and business model innovations, distributed energy resources (DERs) are being deployed in the electric power distribution systems at an unprecedented pace. According to a technical report from Navigant Research [1], the annual installed capacity across the global DER market is expected to grow from 136.4 GW in 2015 to 530.7 GW in 2024.

To fully exploit the benefits of the DERs, the distribution network must be actively managed. To operate the distribution system in an efficient and reliable manner, the distribution system operators typically rely on a set of tools and applications including three-phase power flow, distribution system state estimation, three-phase optimal power flow, distribution system restoration and distribution network reconfiguration. All of these applications require an accurate distribution network and phase connectivity model. Although the network connectivity model is mostly accurate, phasing errors are common [2]. Therefore, an accurate phase identification method is in critical need.

Electric utility companies typically do not have accurate phase connectivity information. Moreover, the phase connectivity of the distribution network changes over time when new customers are connected to the system. With more DERs connected to the power distribution systems, correct phase connectivity data become increasingly important to efficient and reliable operations of power distribution systems. This paper develops an unsupervised machine learning algorithm to identify the phase connectivity of customers based on smart meter data and supervisory control and data acquisition (SCADA) data.

The rest of this paper is organized as follows. Section II introduces the background, provides a comprehensive literature review of the existing methods for phase identification, and clarifies the unique contributions of this paper. Section III presents the proposed phase identification method by clustering of smart meter data. In Section IV, a case study on a Southern California Edison's distribution feeder is conducted to validate the proposed algorithm. The conclusions are stated in Section V.

II. Background and Related Works

A. Background and Problem Definition

To understand the phase identification problem, we first briefly introduce the electric power distribution system. The electric power distribution system is the final portion of the power delivery infrastructure that carries electricity from highly interconnected, high-voltage transmission systems to end-use customers. An illustration of a simple electric distribution system is depicted in Figure 1. The starting point of the distribution system is the distribution substation. In the distribution substation, a step-down transformer lowers the transmission-level voltage (35 to 230 kV) to a medium-level voltage (4 to 35 kV) in the primary distribution circuits [3]. The electric power then flows through the primary feeders and laterals ($L1$-$L5$) to distribution transformers ($T1$-$T8$), which further step down the voltage to low-voltage secondary circuits. The secondary circuits serve end-use customers and operate at 120/240 V three-wire, 120/208 V three-phase, or 277/480 V three-phase. Laterals can be single-phase ($L2$), two-phase, also called "V" phase ($L3$, $L4$), or three-phase ($L1$, $L5$).

The majority of the electric power is supplied by three-phase generators. In balanced conditions, the electric power circuits are 3-phase circuits and the three voltage phasors, V_{an}, V_{bn}, and V_{cn}, differ only in their angles, with 120-degree differences between any pair. Residential customers can be

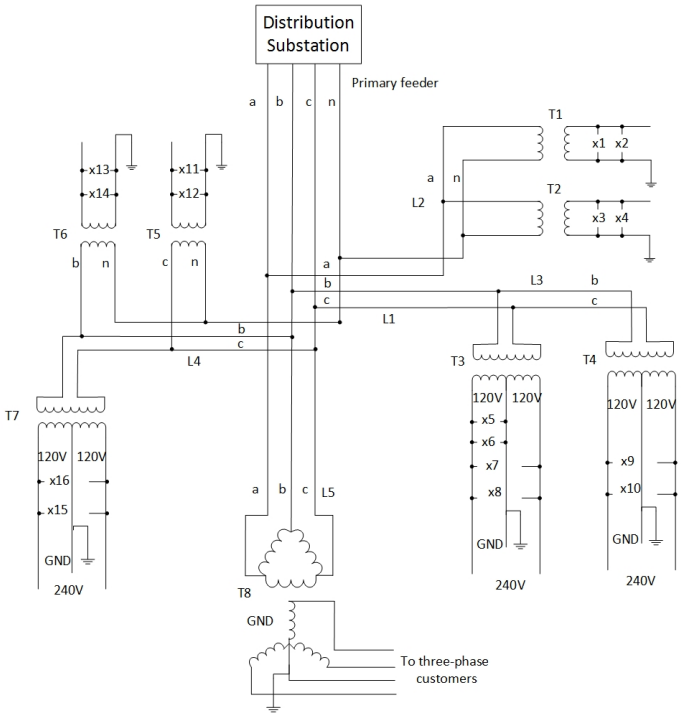

Fig. 1: Illustration of a distribution system. Labels a, b, and c represent the three phases. L stands for a lateral, T stands for a transformer, and x denotes a customer.

served by either a 120/240 V three-wire secondary through a center-tapped transformer (e.g., $T3$, $T4$, $T7$) or a 120 V single-phase secondary through a single-phase transformer (e.g., $T1$, $T2$, $T5$, $T6$). Commercial customers are typically served by a 208 V or 480 V three-phase four-wire secondary through a three-phase transformer (e.g., $T8$).

The phase identification problem is defined as identifying the phase connectivity of each customer and structure in the power distribution network.

B. Related Works and Contributions of This Paper

Very few studies on phase identification have been carried out. The existing methods for solving the phase identification problem can be separated into two general approaches. In the first approach, only smart meter data and SCADA information are assumed to be available [2], [4]–[6]. In the second approach, special equipments such as micro-synchrophasors [7], signal generators and discriminators [8] need to be installed to accurately identify the phase of distribution system customers and/or structures.

In the first approach, 0-1 integer linear programming and correlation-based methods are proposed to solve the phase identification problem. The phase identification problem is formulated as a 0-1 integer linear programming problem where the phase connection of smart meters are treated as binary variables. Tabu search [4] and branch & bound search [5] are used to solve the integer optimization problem. There are two drawbacks associated with the 0-1 integer programming method. The first drawback is its computational complexity. A typical distribution feeder serves 1000 to 3000 customers on average. Therefore, the 0-1 integer programming problem for phase identification has thousands of binary decision variables, which requires daunting computational time. The second drawback is its low tolerance for erroneous and missing measurements. The existing methods only work when there are no unmetered loads or erroneous load measurements.

In correlation-based methods [2], [6], correlation coefficients or R^2 (coefficient of determination) are calculated between the voltage profile of individual smart meters and the voltage profile of the substation on each phase. These correlation coefficients or R^2 are assumed to have the highest value when the customer's phase is correctly labeled. Although correlation-based methods have been shown to be effective in identifying single-phase customers, it is not clear if the method can be successfully applied in the distribution circuits where the majority of the loads are connected to two-phase laterals. In addition, the algorithm may incorrectly label customers on the same single-phase secondary differently.

In the second approach, micro-synchrophasors, signal generators and discriminators are leveraged to accurately identify the phase of each customer. In [7], micro-synchrophasors are deployed at the target bus for phase identification. Micro-synchrophasors can measure voltage phase angles in addition to voltage magnitude. The main idea behind the method is that the correct customer phase label should yield the highest voltage magnitude and phase correlation with the corresponding phase at the substation. The advantage of the micro-synchrophasor approach is that the method is applicable to all types of customer phase connections. In [8], a signal generator is deployed at the distribution substation and signal discriminators are deployed at the target customer sites to accurately identify the phases of smart meters. The disadvantage of the methods in the second approach is the expensive capital and maintenance costs for the additional equipments.

In this paper, an innovative constrained k-means clustering algorithm of smart meter data is proposed to solve the phase identification problem. Instead of directly using the voltage time series data, we propose to first extract unique features from the voltage time series of smart meters. Then we define customer phase constraints by exploiting the known information about line configurations in the network connectivity model. At last, a constrained k-means clustering algorithm is applied to accurately identify the phase connection of each customer.

In light of the existing literature, the unique contributions of this paper are as follows:

1. The proposed phase identification algorithm utilizes the known information about line configurations in the network connectivity model to avoid mislabeling of the customers on the same secondary feeder which can occur in the existing methods.

2. The proposed phase identification algorithm is computationally efficient compared with the 0-1 integer linear programming method and the correlation-based methods.

3. The proposed phase identification algorithm can identify phase connections with high accuracy in distribution circuits where the majority of loads are connected to two-phase laterals.

4. The proposed phase identification algorithm can still determine the phase connections of metered customers when the distribution circuit has some unmetered customers.

III. Phase Identification by Clustering Smart Meter Data

The framework of our proposed phase identification algorithm by clustering smart meter data is illustrated in Figure 2. In the first step, voltage measurements are collected from smart meters and the SCADA system. In the second step, we normalize the customer voltage time series by their standard deviations and apply principal component analysis (PCA) on the normalized time series to extract the top q components. In the third step, we define the constraints in the clustering process by inspecting the network connectivity data. The k-means constrained clustering method is then applied to partition customers into clusters. At last, we identify the phase of each cluster by solving a minimization problem. The rest of this section is divided into three parts. First, we briefly review the methods in clustering of time series data. Second, the k-means constrained clustering algorithm for smart meter data is presented. Third, the algorithm for identifying the phase of each cluster is introduced.

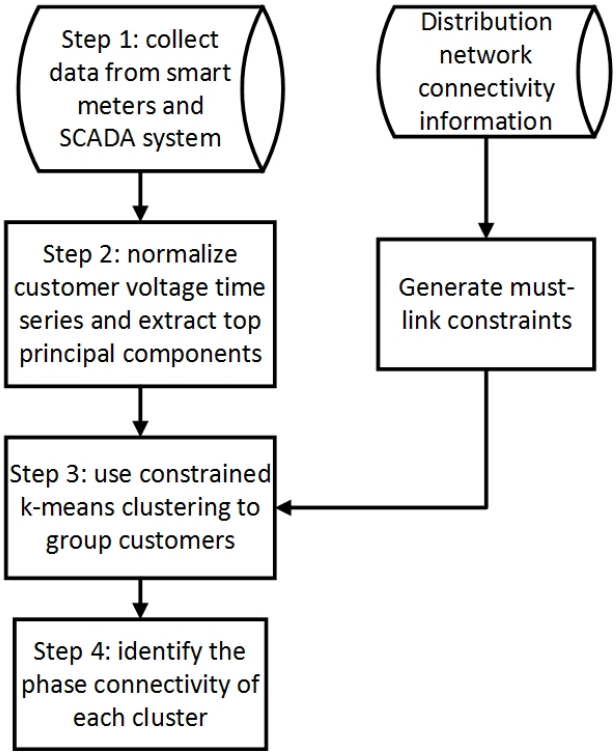

Fig. 2: Diagram of the phase identification procedure.

A. Brief Review of Clustering Time Series Data

The goal of clustering is to identify the structure in an unlabeled dataset by objectively organizing data into homogeneous groups such that the objects in the same group are more similar to each other than those in different groups [9]. Various algorithms have been developed to cluster time series data. One of the widely used clustering algorithms is k-means, in which the objects are divided into k clusters so that the within-cluster sum of squares is minimized. Though typically it is not practical to find the minimal sum of squares among all possible partitions, many algorithms have been proposed to find local optimal solutions [10].

Almost all clustering algorithms require a similarity or distance function. There are many different types of distance functions. We only consider two of them here. The first one is Euclidean distance. If a_i and a_j are two p-dimensional time series, then their Euclidean distance is defined by

$$d_E = \sqrt{\sum_{k=1}^{p}(a_{ik} - a_{jk})^2} \tag{1}$$

Another type of distance function is related to Pearson's correlation coefficient. For two p-dimensional time series a_i and a_j, their Pearson's correlation factor is defined by

$$cc = \frac{\sum_{k=1}^{p}(a_{ik} - \mu_i)(a_{jk} - \mu_j)}{s_i s_j} \tag{2}$$

where μ_i and μ_j are the mean values of a_i and a_j, and $s_i = \sqrt{\sum_{k=1}^{p}(a_{ik} - \mu_i)^2}$ [9]. Then the distance between a_i and a_j can be defined based on cc as $d_1 = 1 - cc$ or $d_2 = (\frac{1-cc}{1+cc})^{\beta}$, $(\beta > 0)$ [11].

Smart meter time series data are high-dimensional. It is not desirable to work with high-dimensional noisy raw data in practice [9]. Therefore, we adopt a feature-based clustering method for the phase identification problem. Drawing features from data often requires expert knowledge of the data, but in the phase identification problem, little knowledge is known on what features are important. PCA is a useful tool to reduce the data dimension and extract key features hidden in the time series data. PCA transforms a dataset into a new set of uncorrelated variables called principal components (PCs). PCs are ordered such that the first component retains the most of the variation in the original variables, the second component retains the second most of the variation, and so on [12]. In this paper, PCA is used to select the most important features of the voltage time series data by picking the first q components. Euclidean distance in the chosen principal components' space will be used as the distance metric in the subsequent clustering process.

B. Clustering of Smart Meter Data with Constraints

The intuition behind identifying phase connectivity through clustering of voltage time series data is that the distribution system is typically operated in an unbalanced manner. The unbalanced impedances and electric loads on three phases lead to unbalanced line currents and voltages [13]. This implies that

the trajectory of voltage time series of customers with the same phase connectivity will have more similar behavior than those with different phase connectivity. As mentioned in Section III-A, instead of working directly with the raw voltage data, a feature-based clustering approach is adopted with features extracted from the voltage time series by PCA. Preprocessing including normalization and centering of the raw voltage data is conducted before applying PCA. We will show in the case study in Section IV that a small number of features can yield very accurate clustering results.

The goal of clustering the voltage data from smart meters is to identify distinct groups of customers such that all customers in the same group have the same phase connectivity. Using the distribution feeder shown in Figure 1 as an example, customers $x7$, $x8$, $x9$, $x10$, $x15$, and $x16$ are all connected to phase BC through a three-wire system (120/240 V) and they should be clustered into the same group. Similarly, consumers $x1$, $x2$, $x3$, and $x4$ should also be in one cluster because they are all connected to phase A and have the same voltage level (120 V). Before applying the clustering algorithm, we first separate customers based on their service voltage levels (120 V, 120/240 V, 208 V, 277 V, 480 V). These voltage levels can be easily identified by inspecting the voltage magnitude data from smart meters. The algorithm proposed in this paper aims at clustering customers of the same voltage level. For example, meters of 120/240 V three-wire service have 6 possible phase connections: AB, BC, CA, AN, BN, and CN; meters of 120 V single phase service have 3 possible phase connections: AN, BN, and CN.

Various studies have been carried out to improve clustering/learning performances by utilizing constraints from background knowledge [14]–[17]. In [14], two kinds of hard constraints are introduced: *must-link* constraints and *cannot-link* constraints. Must-link constraints specify that two data points have to be in the same cluster; cannot-link constraints specify that two data points cannot be in the same cluster. The constraints for the phase identification problem can be formed based on the network connectivity information, which is typically available for power distribution systems. The network connectivity information includes line segment configurations and the connectivity between customers, distribution transformers, laterals, and primary feeders. If two customers are connected to the same secondary laterals and have the same voltage level, then they must have the same phase connectivity and should be linked together in the clustering process. For example, in Figure 1, customers $x7$, $x8$, $x9$, and $x10$ are all connected to the same lateral $L3$, and receive power through a three-wire (120/240 V) configuration. Therefore, these customers should be grouped into the same cluster. However, customers $x7$ and $x15$ should not be linked to each other because they are connected to different laterals.

A scheme is introduced in [15] for constrained k-means clustering. It is similar to the standard k-means clustering algorithm except that in the constrained clustering algorithm, each data point is assigned to the closest cluster such that it does not violate the constraints. The phase identification problem has must-link constraints where certain data points must be in the same cluster. We first put customers on the same laterals into a subset. Then an augmented k-means clustering algorithm is performed to the subsets themselves to obtain the full partition. Let $D = D_1 \bigcup D_2 \bigcup ... \bigcup D_n$ be the whole dataset, and $D_1, ..., D_n$ are the subsets in which every data point is linked together by the constraints. If a data point is not linked to any other data point, then it forms a subset in D itself. The constrained k-means clustering algorithm for phase identification is described in Algorithm 1, which is a modification of the scheme in [15]. As mentioned in Section III-A, it is difficult to find the optimal result(s) by k-means clustering. To get a relatively good clustering result in our approach, the clustering algorithm is performed multiple times with different sets of random initial cluster centers. The clustering result with the smallest sum of squared distances is selected in the end.

Algorithm 1 Constrained k-means clustering algorithm

1: **procedure** CON-K-MEANS($D = D_1 \bigcup D_2 \bigcup ... \bigcup D_n$)
2: Choose data points randomly from D as the initial cluster centers $C_1, ..., C_k$.
3: Calculate each subset D_i's distance to each cluster. The distance is defined as the sum of squared distances of all the data points in D_i with the cluster center.
4: Assign each subset to the cluster that has the minimum summed distance.
5: For each cluster C_i, update its center by averaging all the data points that have been assigned to it.
6: Iterate between (3) and (5) until convergence.
7: **return** $\{C_1, ..., C_k\}$.
8: **end procedure**

C. Identify the Phase Connectivity of Each Cluster

Once the customers are clustered as described in Section III-B, the next and last step is to identify the phase connectivity of each cluster. Since the customers in the same cluster should have the same phase connection, we can identify the phase of each cluster by picking a small number of customers from that cluster and identify their phase connectivity. This is a huge workload reduction compared with performing phase identification algorithms on every single customer. One may identify the phase of these few customers by micro-synchrophasors, signal generators and discriminators as in [7], [8].

However, to further reduce the computational workload, and to save the expense of equipments used in [7], [8], we can identify the phase of each cluster by a one-to-one matching between the set of clusters and the set of possible phase connections. The one-to-one matching can be found by solving the following minimization problem. Suppose there are k clusters to be identified with centers $C_1, ..., C_k$, and there are k substation voltage time series on the k possible phases. The k substation voltage series are centered and normalized by their standard deviations, and then projected onto the chosen principal components' space used for clustering. Let $V_1, ..., V_k$ be

the coordinates of the k voltage series in the chosen principal components' space, and let $f : \{C_1, ..., C_k\} \rightarrow \{V_1, ..., V_k\}$ be an unknown bijection between the cluster set and the substation voltage set. The solution of the minimization in (3) is the one-to-one matching for phase identification. The phase of each cluster's paired voltage data is the cluster's identified phase.

$$\underset{\forall \text{ bijection } f:\{C_1,...,C_k\}\rightarrow\{V_1,...,V_k\}}{\arg\min} \sum_{i=1}^{k} d_E(C_i, f(C_i))^2 \qquad (3)$$

Here $d_E(C_i, f(C_i))$ is the Euclidean distance between C_i and $f(C_i)$. The minimization can be solved by exhaustive search, because there are only $k!$ possible bijections, where k is small (e.g., $k = 3$ at 120/240 V level).

Compared to the load matching approach in [5], which assumes aggregated electricity consumption of all customers matches that of the substation, our proposed method is less sensitive to the presence of unmetered customers.

IV. Case Study: Southern California Edison Distribution Feeder

In this section, the proposed phase identification method is validated through a case study of a distribution feeder in Southern California Edison's service territory. The results show that the constrained k-means clustering algorithm yields highly accurate phase connectivity on a typical distribution feeder.

A. Description of Datasets and Preprocessing of Data

The distribution feeder used for case study is a 12.47 kV network with a peak load of about 5.2 MW. The feeder serves about 1500 customers. The majority of the customers are residential customers.

The raw data collected to test the phase identification algorithm include: 1) hourly smart meter readings of voltages; 2) feeder line-to-line voltage readings of three phases from the SCADA system; 3) network connectivity of the distribution system. The number of a smart meter's readings varies by month. In months with 30 days, there are 720 readings (yielding measurement vectors of dimension 720), while months with 31 days have 744 reading hours. The SCADA system only records new feeder measurements when the difference between the new measurement and the previous measurement exceeds certain threshold. For example, the threshold setting for the line-to-line voltage is 0.02 kV. At last, to evaluate the accuracy of the proposed phase identification method, the correct phase connectivity of each meter is also gathered to serve as the ground truth.

Since the SCADA readings are recorded at nonuniform timestamps, linear interpolation is used to create a new set of voltages that have the same timestamps as the smart meter readings. All the readings are centered and normalized by their standard deviations. PCA and k-means clustering are performed on the readings of the same time period with the same timestamps. The timestamps are chosen such that most meters have a complete set of measurements. A smart meter is removed from the case study if it has missing readings at the chosen timestamps in the study period. In the testing distribution feeder, most of the customers are served by a three-wire system (120/240 V) based on the smart meter voltage levels, and all of them are connected to phases AB, BC, or CA. A few customers are served by three-phase laterals and there is no need to perform phase identification for these customers. Less than 1% of the customers are served by two-wire single-phase systems (120 V). Due to the small number of datasets, they are removed from the clustering process and their phase connectivity can be identified using methods introduced in [7], [8].

After preprocessing the test data, about 1500 customers/meters need to be clustered into 3 groups: phase AB, phase BC, and phase CA. PCA is conducted on the preprocessed time series data. Only the first two principal components are used to calculate Euclidean distances among customers. Based on the simulation results, including additional principal components does not further improve the performance of the phase identification results. The phase of each cluster is identified by finding the bijection described in Section III-C. In this case, the bijection is between 3 clusters and the substation voltages of phases AB, BC, and CA.

B. Clustering Results

TABLE I: Clustering Result

Unconstrained Clustering Results of August 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	674	92.58%	87.55%
2	BC	518	87.64%	
3	CA	246	73.58%	

Constrained Clustering Results of August 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	636	98.27%	90.40%
2	BC	560	87.68%	
3	CA	242	76.03%	

Unconstrained Clustering Results of September 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	678	93.36%	93.12
2	BC	547	93.60%	
3	CA	244	91.39%	

Constrained Clustering Results of September 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	645	98.29%	97.28%
2	BC	559	97.67%	
3	CA	265	93.96%	

Unconstrained Clustering Results of October 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	662	95.02%	93.09%
2	BC	531	93.60%	
3	CA	254	87.01%	

Constrained Clustering Results of October 2015

Cluster	Identified Phase	Number of Meters	Accuracy	Overall Accuracy
1	AB	630	99.84%	97.86%
2	BC	550	98.36%	
3	CA	267	92.13%	

Three months of SCADA, smart meter, and network connectivity data are collected from August 1, 2015 to October 31, 2015. 1438 smart meters' data are available in August. According to the ground truth, 629 of them are connected to phase AB laterals, 557 of them are connected to phase

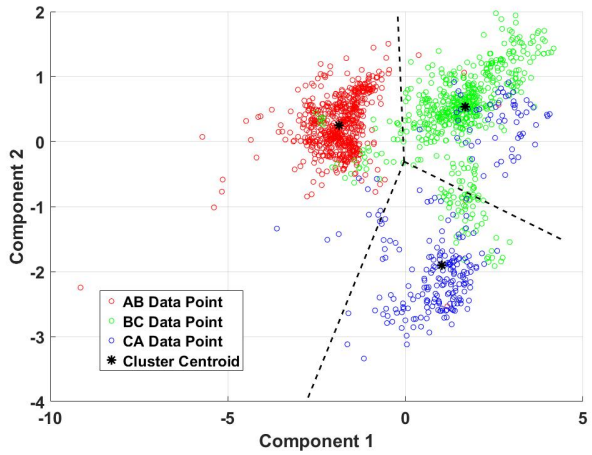

Fig. 3: Principal components of August voltage time series data.

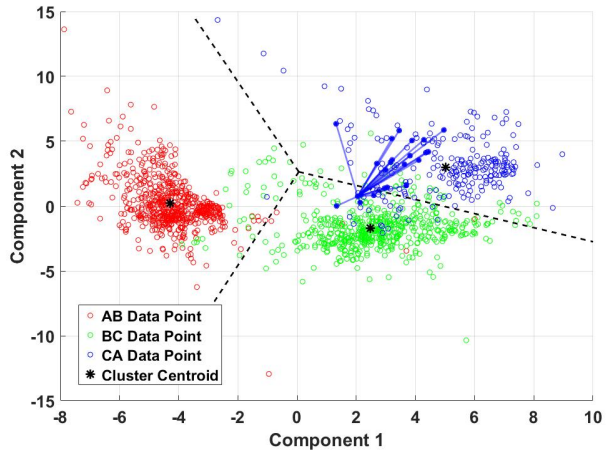

Fig. 4: Principal components of October voltage time series data.

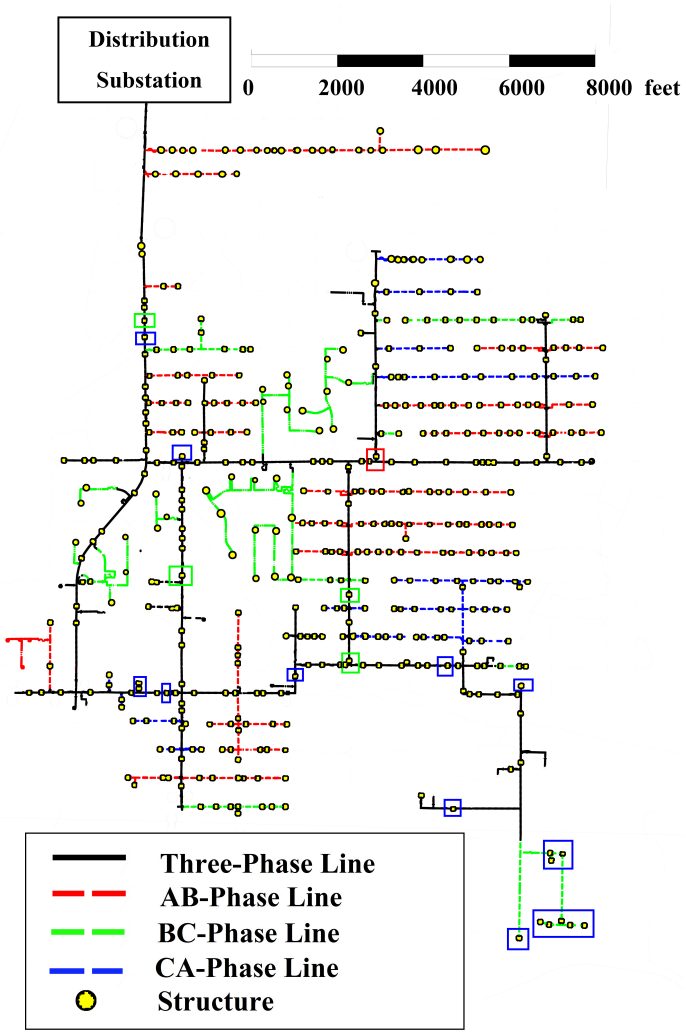

Fig. 5: Phase identification results.

BC laterals, and 252 of them are connected to phase CA laterals. In September, 1469 smart meters' data are available. According to the ground truth, 638 of them are connected to phase AB laterals, 571 of them are connected to phase BC laterals, and 260 of them are connected to phase CA laterals. In October, 1447 smart meters' data are available. According to the ground truth, 633 of them are connected to phase AB laterals, 562 of them are connected to phase BC laterals, and 252 of them are connected to phase CA laterals.

The clustering and phase identification results are shown in Table I. These results can be interpreted as follows. The clustering and phase identification algorithms group the smart meters into three clusters. The phase identified for each cluster is listed in the identified phase column. If a meter is assigned to a cluster whose identified phase is the same as the meter's actual phase, then it is assigned to the correct cluster. The accuracy column shows the percentage of correct assignments in each cluster and the overall accuracy column shows the overall accuracy of the phase identification algorithm.

Table I shows that the phase identification algorithm of both unconstrained and constrained clustering achieve at least 90% overall accuracy in September and October. In addition, in all months, the constrained clustering algorithm yields a higher accuracy than the unconstrained k-means clustering algorithm. The constrained clustering outperforms the unconstrained clustering by letting must-link constraints pull a linked meter back to the correct cluster when it is near the boundary of two clusters.

Figure 3 and 4 show the distributions of two months' voltage data points in the space of the first two principal components. Dashed lines are the boundaries of Voronoi cells associated with cluster centers derived from the constrained clustering algorithm. Figure 4 also shows an example of how the constrained clustering algorithm improves the accuracy. In Figure 4, a set of blue data points grouped by must-link constraints are connected by solid lines. Although this set of data points are separated by a boundary, they are closer to the CA cluster as a whole. Therefore, they are assigned to the

CA cluster, which is the correct phase. Without these must-link constraints, some of the data points will be assigned to the BC cluster, which is incorrect. Figure 3 and 4 show that data points of different phases are separated in the space of the first two principal components. However, there are more data points of phase BC and CA overlapped in Figure 3 than Figure 4. As a result, the overall accuracy of phase BC and CA are lower when using data from August, compared with October.

Figure 5 shows the clustering results on the distribution circuit map based on the smart meter data of October 2015. In Figure 5, each line is colored according to its actual phase. Each structure (e.g., transformer) is represented by a small dot. The three-phase black lines are primary feeder lines. Structures can be connected to primary feeder lines through a three-wire (120/240 V) system, so they can be connected to phases of AB, BC, and CA. A colored rectangle is overlayed on top of a structure if it is assigned to a wrong cluster. The color of the rectangular shows the identified phase of the cluster. Note that the number of structures is smaller than the number of smart meters/customers as a distribution transformer typically serves several customers.

The results above show that the constrained k-means clustering algorithm groups the meters by phase at high accuracy, and the identification method correctly identifies the phase of each cluster, in a circuit where the majority of customers are connected to two-phase laterals.

The proposed algorithm is computationally more efficient than the integer linear programming method. The running time of the proposed algorithm is the sum of the running time of the PCA step and the k-means clustering step. The running time of the PCA is $O(p^2m + p^3)$ [18], and the running time of Lloyd's algorithm for k-means clustering is given by as $O(mkqi)$. Here p is the number of dimensions of the raw time series data, m is the number of data points (i.e., the number of meters), k is the number of clusters, q is the number of principal components used in clustering, and i is the number of algorithm iterations. In the case study, the typical value for i is less than 10. Therefore, the total running time of the proposed algorithm increases linearly with m. On the other hand, the running time of branch and bound search, which solves the integer linear programming problem, is not bounded by a polynomial function of m [19].

V. Conclusions and Future Work

An innovative distribution system phase identification algorithm using constrained k-means clustering of smart meter data is proposed in this paper. The proposed algorithm leverages the network connectivity information to avoid mislabeling of customers on the same secondary feeder. Utilizing only the smart meter and SCADA information, the proposed algorithm is not only computationally efficient but also yields high accuracy. A real-world distribution feeder is used as a test case to validate the proposed algorithm. The case study results show that the constrained k-means clustering algorithm outperforms the unconstrained algorithm. The overall accuracy of the proposed algorithm is at least 90%.

Table I shows that this algorithm performs better during some months than others. Future research is needed to determine over which time periods the phase identification algorithm performs best. In addition, it is desirable to develop algorithms that not only perform phase identification but also estimate the confidence level of clustering result for each individual meter.

Acknowledgment

The authors would like to thank Austen D'Lima and Marshall Parsons from Southern California Edison for fruitful discussions and supplying AMI data, network connectivity data, and field validation data.

References

[1] "Distributed energy resources global forecast," Navigant Research, Tech. Rep., 2015.

[2] T. A. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 651–658, Jun. 2013.

[3] T. A. Short, *Electric Power Distribution Handbook*, 2nd ed. CRC press, 2014.

[4] M. Dilek, "Integrated design of electrical distribution systems: phase balancing and phase prediction case studies," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2001.

[5] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavlovski, S. Hoy, and J. R. Kalagnanam, "Phase identification in smart grids," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE, 2011, pp. 25–30.

[6] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase LV distribution network," in *Universities Power Engineering Conference (AUPEC), 2012 22nd Australasian*. IEEE, 2012, pp. 1–7.

[7] M. Wen, R. Arghandeh, A. Meier, K. Poolla, and V. Li, "Phase identification in distribution networks with micro-synchrophasors," in *2015 Power and Energy Society General Meeting*. IEEE, 2015, pp. 1–5.

[8] K. Caird, "Meter phase identification," Jul. 2010, U.S. Patent App. 12/345,702. [Online]. Available: http://www.google.com/patents/US20100164473

[9] T. W. Liao, "Clustering of time series data-a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[10] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, Jan. 1979.

[11] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, and P. Boesiger, "A new correlation-based fuzzy logic clustering algorithm for FMRI," *Magnetic Resonance in Medicine*, vol. 40, no. 2, pp. 249–260, Aug. 1998.

[12] I. Jolliffe, *Principal Component Analysis*, 2nd ed. Wiley Online Library, 2002.

[13] W. H. Kersting, *Distribution System Modeling and Analysis*, 3rd ed. CRC press, 2012.

[14] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," *AAAI/IAAI*, vol. 1097, 2000.

[15] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.

[16] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, pp. 81–88.

[17] T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann, "Learning with constrained and unlabelled data," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 731–738.

[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. JHU Press, 2012, vol. 3.
[19] A. Schrijver, *Theory of Linear and Integer Programming*, 1st ed. John Wiley & Sons, 1998.