Voltage Correlations in Smart Meter Data

Rajendu Mitra, Ramachandra Kota, Sambaran Bandyopadhyay, Vijay Arya
IBM Research
{rajendum, rama.chandra, sambandy, vijay.arya}@in.ibm.com

Brian Sullivan, Richard Mueller, Heather Storey, Gerard Labut
DTE Energy, USA
{sullivanbj, muellerrj, storeyh, labutg}@dteenergy.com

ABSTRACT

The connectivity model of a power distribution network can easily become outdated due to system changes occurring in the field. Maintaining and sustaining an accurate connectivity model is a key challenge for distribution utilities worldwide. This work shows that voltage time series measurements collected from customer smart meters exhibit correlations that are consistent with the hierarchical structure of the distribution network. These correlations may be leveraged to cluster customers based on common ancestry and help verify and correct an existing connectivity model. Additionally, customers may be clustered in combination with voltage data from circuit metering points, spatial data from the geographical information system, and any existing but partially accurate connectivity model to infer customer to transformer and phase connectivity relationships with high accuracy. We report analysis and validation results based on data collected from multiple feeders of a large electric distribution network in North America. To the best of our knowledge, this is the first large scale measurement study of customer voltage data and its use in inferring network connectivity information.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*; J.m [**Computer Applications**]: Miscellaneous

General Terms

Algorithms, Design, Measurement, Experimentation

Keywords

Power Distribution Grids; Voltage Time Series; Topology Inference; Data Mining; Clustering

1. INTRODUCTION

The connectivity model (CM) of the physical distribution network specifies how the devices, assets, and customers are interconnected downstream of a distribution substation. For example, which customer is powered by which distribution transformer, which customer is powered by which phase of the feeder, and so on. CM essentially provides a coarse-grained view of the network topology. A common problem faced by distribution utilities worldwide is an inaccurate or unknown CM of their network when compared to the actual connectivity that exists on the field. The CM may not always be updated or tracked based on changes made by field crews and its accuracy deteriorates over time due to maintenance, repairs, and restoration activities following faults or outages. Moreover during large scale outages, there is often a trade-off between expediting restoration versus tracking changes to the distribution network.

While CM is foundational to planning, operations, and maintenance of distribution networks, the key factors driving utilities to improve its accuracy are effectively faster restoration and the ability to accurately communicate with impacted customers during outages. The annual cost of power interruptions in US is estimated to be $79B with 106 ± 54 outage minutes per customer. Interruptions in electric service occur from time to time due to a number of reasons including storms, aging assets, excess loading from heat waves, and other system disturbances. Any analysis following a fault in the distribution system uses the CM to identify the root causes and determine the appropriate course of action. An accurate CM minimizes the diagnostic time and the time spent by crew in the field, leading to reduced outage minutes and improved system availability.

During outages, utilities seek to inform customers about the status of restoration and the expected downtimes. The CM is required to localise customers downstream of a faulted device and to map each fault with the right set of outaged customers for communication. An inaccurate CM increases the risk of erroneous communication and limits a utility's ability to provide customized and timely information to their customers, which may negatively impact customer relations. Additionally, an accurate CM is required to localise losses, estimate loading at unmetered points such as distribution transformers, and ensure a balanced infusion of energy back into the distribution grid when customers have behind-the-meter resources such as distributed generation and storage/electric vehicles [4, 6, 9, 10].

Existing techniques to maintain an accurate CM include the use of manual field inspections and power line communications (PLC). Manual inspections are expensive and unsustainable as field configurations change over time and therefore these need to be repeated periodically. The PLC

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

ACM 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2788594 .

approach requires both customer and grid meters to be able to read and write signals onto the power-line and is capital intensive. Additionally, challenges arise as signals may not propagate over long distances or across assets.

Leading utilities are undertaking initiatives to modernize their information management and data analytics capabilities in order to realize the benefits of smart grid deployments. This work focuses on the use of data already available from a distribution network to infer its connectivity model. Our approach to infer connectivity relies on clustering customers using their voltage data. The computed customer clusters may be used to identify inconsistencies in the existing CM or correct an existing CM that may be partially accurate. We observe that voltage time series measurements from customer smart meters exhibit hierarchical correlations which may be exploited to partition customers based on common phase[1], distribution transformer, and feeder. Customers may also be clustered in combination with voltage data from circuit metering points (SCADA) and spatial data from the GIS to verify and correct errors in customer to transformer and phase mapping, which are generally prone to inaccuracies. We report validation results based on the analysis of data collected from multiple feeders of a large distribution network in North America. The data includes average RMS values of voltage recorded once every 5 minutes from more than 10K customers across 5 feeders and 2 substations for about 2 months. The performance of techniques is estimated by comparing the computed connectivity relationships with the ground truth available from existing database and results from field verification. We view the main contributions of this work as follows:

1. Using smart meter measurements, we show that customer voltage data exhibits hierarchical correlations which tend to be consistent with the hierarchical connectivity relationships between customers, distribution transformers, phases, feeders, and substation transformers of a distribution network. Towards this, we transform the voltage measurements into binary fluctuations and conduct different clustering experiments by considering customers within and across different feeders and substations. We show that customers can be partitioned based on common phases and feeders with high accuracy. We point out settings where phase correlations are stronger when compared to feeder correlations.

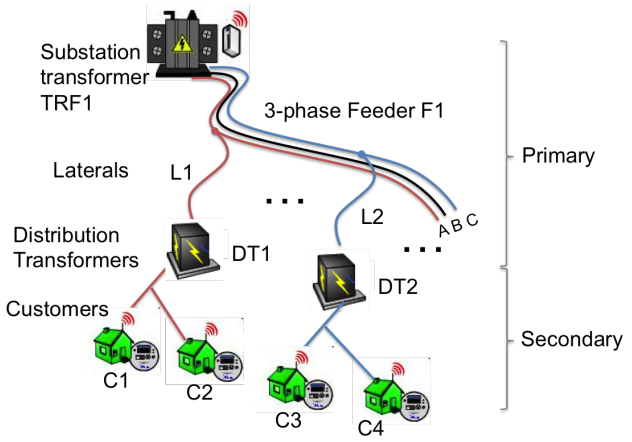
2. We compare different approaches to infer customer to transformer mapping by clustering customers under a feeder, its individual branches, and within spatial subclusters formed from GIS data. The clustering solutions are compared with the currently available mapping in the database which is known to be accurate. Our results indicate that clustering customers under each branch of the feeder introduces constraints that help improve the accuracy of clustering.

3. To infer customer to phase mapping, customer voltage measurements are clustered in combination with the per-phase feeder measurements obtained from SCADA data which act as centroids of an initial solution. We compare the computed solutions with the mapping that was available both before and after manual field inspections. Our results indicate that the proposed approach is able to correct most of the errors identified during field inspections and may be used as a low cost alternative to help maintain an accurate phase connectivity model.

[1]The concept of *phase* and the structure of the distribution network are explained in Section 2

The rest of the paper is organized as follows. Section 2 explains the topology of the distribution network and the connectivity model. Section 3 summarizes the voltage and spatial data used in this work and our observations on various correlations, while Section 4 describes our clustering methodology. Section 5 presents our analysis of the hierarchical correlations. Sections 6 and 7 describe our techniques and results for inferring customer to transformer and phase connectivity using clustering. Section 8 presents related work and Section 9 concludes.

2. DISTRIBUTION NETWORK, CONNECTIVITY MODEL, AND ERRORS



Substation Transformer	Feeder	Distribution Transformer	Cust.	Phase
TRF1	F1	DT1	C1	A
TRF1	F1	DT1	C2	A
TRF1	F1	DT2	C3	C
TRF1	F1	DT2	C4	C

Figure 1: Simplified view of a distribution network and the mapping between customers and assets

Distribution System. The power delivery infrastructure consists of generation, transmission, and distribution systems. Electric power is generated in large power plants as 3-phase AC voltage and reaches distribution via a transmission system. The distribution system starts from the distribution substation and consists of primary and secondary networks. The primary network consists of 3-phase *feeders* which carry power at medium voltage from the *substation transformers* (TRFs) to the *distribution transformers* (DTs). Each TRF may power multiple feeders. Depending on the type and density of customers, a feeder may carry power at different voltages such as 33kV, 13.2kV, and 4.8kV and serve a few hundred distribution transformers. The branches or line segments of a feeder are also known as *laterals*. The secondary network carries power from the distribution transformers to the customers at low voltage (e.g. 120V).

Table 1: Information about the feeder circuits analyzed from the distribution network

Substation, Substation transformer	Feeder circuit	Voltage fed	Length (Miles)	Residential	Non-residential	1-phase transformers	1-phase customers	Missing data
SUB-I, TRF-I	α	13.2kV	34	80%	20%	221	1773	12%
SUB-II, TRF-I	β_1	13.2kV	38	97%	3%	342	2178	1%
SUB-II, TRF-I	β_2	13.2kV	39	91%	9%	306	2363	1%
SUB-II, TRF-II	γ_1	13.2kV	38	92%	8%	363	2497	1%
SUB-II, TRF-II	γ_2	13.2kV	30	94%	6%	281	1561	1%

A 3-phase feeder consists of three transmission lines, usually labelled as A, B and C, which carry AC power with their voltage waveforms shifted by 120°. A DT receives power by tapping onto one of the 3 phases of a feeder and is generally single-phase (1-phase). On average a DT might serve about 8 single-phase residential customers. A few DTs that serve larger loads such as super-markets or office buildings may be 2 or 3-phase. Similarly a feeder branch or lateral may be single, two, or three-phase. [10].

Connectivity Model. The CM of a distribution network specifies the mapping between various assets and customers downstream of a substation. In particular, which customer is powered by which DT, which DT is powered by which feeder and phase, which feeder is powered by which TRF, and so on. For example, the CM of the distribution system in Fig. 1 records that customers C1, C2 are powered by distribution transformer DT1, while C3, C4 are powered by DT2. DT1 is powered by phase A of feeder F1, while DT2 is powered by phase C of F1. Feeder F1 in turn is powered by TRF1. In the above example, since the distribution transformers are single-phase, customer phase is same as transformer phase. The CM may also record other hierarchical relationships such as the mapping between a lateral and the set of transformers that it powers.

Errors. The most common errors in the CM are related to the mapping between customers, DTs, and phases of the feeder. For example, the phase of a customer or its distribution transformer may be recorded incorrectly or missing. Similarly the DT that powers a customer may have been inaccurately recorded. Other types of errors include the mapping of customers to feeders wherein some of the customers under a feeder are recorded as being powered by another feeder. We study these mappings in subsequent sections.

3. DATA CHARACTERISTICS

Power measurements in a distribution network are generally available from customer smart meters and on-grid sensors. The primary network is monitored using a SCADA (supervisory control and data acquisition) system that records feeder level measurements (per-phase voltage, current, active and reactive power, etc.) close to the substation. customer smart meters generally record periodic measurements of load (kWh) and voltage (RMS value) over small time intervals of 5-30 min as setup by the utility.

In this work we use measurements from five anonymized feeder circuits belonging to two substations of a North American distribution network. The mapping between substations, TRFs and feeders is showing in Table 1. Each circuit is fed at 13.2kV and serves more than 2K residential, commercial and industrial customers.

Low-voltage data from smart meters. We analyze two months of voltage measurements from customers which are powered by single-phase distribution transformers. The measurements have average RMS values of voltage in the 120V range. These are recorded once every 5 minutes up to a precision of one decimal place. The measurements also have missing values.

Medium-voltage data from the SCADA system. The feeder circuits are instrumented with SCADA field devices which monitor the feeders close to the substation transformer and periodically report per-phase measurements of voltage, current, and power. We consider two months of voltage measurements from these devices. The measurements are averaged over 5 minute intervals and normalized to 120V range to compare them against the low-voltage data dfrom the customers.

Spatial Data from GIS. Anonymized geospatial locations of customers and DTs are available and used to subcluster customers and transformers under a feeder based on geographical proximity. This is utilized in the inference of customer to DT mapping.

Customer to DT mapping (Ground Truth). The existing CM holds the mapping between customers and the DTs. While this mapping is known to have good accuracy for the 5 chosen circuits, it may not be 100% accurate. We use this mapping to estimate the performance of clustering algorithms.

Customer to phase mapping (Ground Truth). The transformer phase is available for all 1-phase distribution transformers (customer phase is same as transformer phase for 1-phase transformers). The phase of overhead (pole top) transformers is available before and after manual field inspections. In case of overhead distribution, field inspections are conducted by linemen using a combination of visual inspections and specialized devices. The devices are reserved to check the phase at certain operating points in the feeder while walking down the line segments and visually inspecting the phase down to individual transformers or customers. The voltage data analyzed in this work was collected before commencing field inspections. The phasing information available during data collection and after field inspections is used to validate algorithms and also identify errors in existing phasing information.

DT to lateral mapping. Accurate mapping from distribution transformers to laterals, i.e. the set of customers powered by each branch of the feeder, is available in the existing CM for each of the circuits. This mapping is used to introduce spatial constraints in the inference of customer to DT mapping.

3.1 Observations on Voltage Data

Fig. 2 shows benchmark plots that give an overview of the voltage measurements obtained from customer smart meters and the SCADA system. Fig. 2 (left) plots the voltage measurements of 3 random customers on different phases

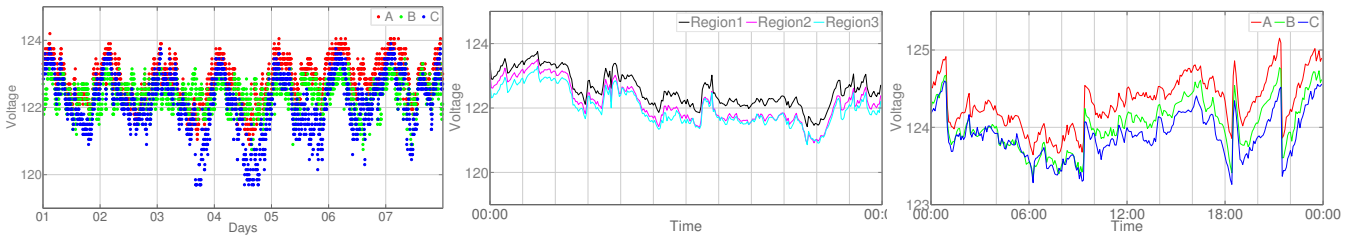


Figure 2: Voltage data from one of the feeders and its customers. (left):Voltage of 3 customers on phases A, B, and C over 7 days, (center): Average voltage of customers in 3 different sections of a feeder circuit over a day, (right): Per-phase feeder measurements taken close to substation over a day.

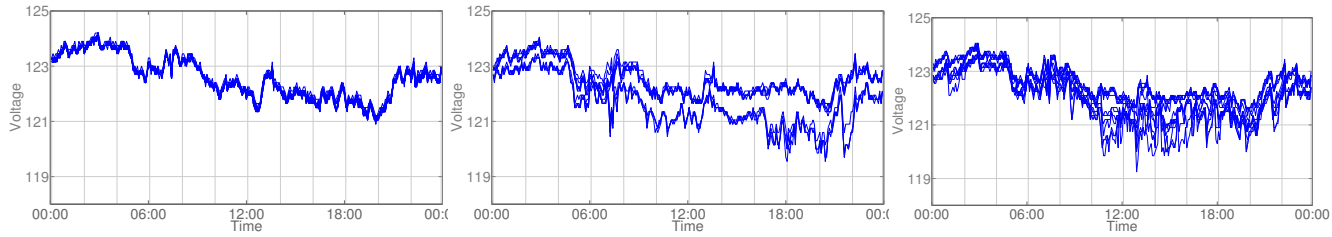


Figure 3: Different types of voltage measurements. Measurements of all customers under each DT.

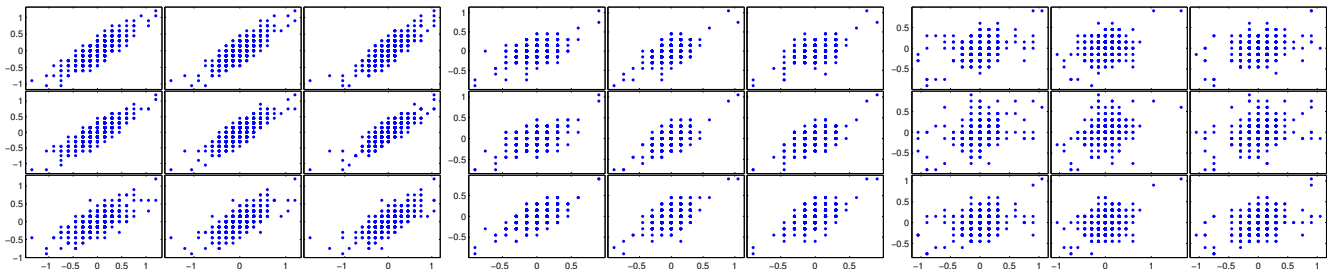


Figure 4: Correlations of customer voltage fluctuations using scatter plot– (left): Two different sets of customers under same transformer (hence same phase) (center): Two sets of customers under different transformers but same phase. (right): Two sets of customers under different phases (hence different transformers).

and under different distribution transformers over a 7 day period. We observe the diurnal cycles of voltage based on energy consumed each day and differences in voltage across customers based on phase. In order to study the voltage observed in different regions of a feeder circuit, we divided the circuit into three sections based on deployed reclosers (circuit breakers). Fig. 2(center) plots the average voltage of customers in each of these regions. Region 1 corresponds to customers which are closer to the distribution substation. These customers observe a slightly higher voltage irrespective of phase. Customers in regions 2 and 3 observe a small voltage drop as they are located further away from the substation. The plot demonstrates that customer voltage also depends on a number of factors other than phases or distribution transformers. Fig. 2(right) plots the per-phase voltage measurements observed at the feeder normalized to 120V range for a day. We see that the voltages on the 3 phases vary together on larger time scales based on aggregate load while each phase shows distinct variations at smaller time scales. Although voltages on the 3 phases differ, the feeder voltage measurements showed low imbalances on average across all days.

Fig. 3 shows voltage plots of customers under different distribution transformers. Fig. 3 (left) shows an example where voltages of all customers under a transformer are very similar and vary together. Fig. 3 (center) shows another transformer where the customer voltages vary within two groups. This may occur due to variable loading or distances of customers from their transformers. Fig. 3 (right) shows a transformer where the voltages of customers show high variation. The plots show that deciphering customer voltage may be challenging as it depends on several factors in the electric network.

Fig. 4 shows the benchmark correlations between voltage fluctuations of customers under one feeder using a matrix of scatter plots. Fig. 4 (left) shows correlations between two different sets of customers which are powered by the same single-phase transformer (hence same phase as well). Each scatterplot shows correlations between a pair of customers. We observe that voltage fluctuations of customers under the same transformer show high correlation and points lie closer to the X = Y line. Fig. 4 (center) shows correlations between two sets of customers powered by different transformers under the same phase of the feeder. We observe that correlations still exist but become weaker in comparison to the same transformer case. Lastly, Fig. 4(right) shows correlations between two sets of customers powered by different phases of the feeder (hence different transformers as well). We see that these customers are even less correlated and points are spread over the X = Y plane. In section 5, we also consider correlations of customers across different feeders.

4. DISTANCE METRICS & CLUSTERING

Our approach to infer connectivity leverages the fact that the voltage variations observed by customers powered by the same DT, phase, or feeder tend to be more "similar". However, there are different measures of similarity and different approaches to compare voltage measurements. In particular, we have studied 3 distance metrics and 3 data transformations which are described below.

4.1 Voltage data transformations

Let $v_{i,t}$ denote the raw RMS voltage measurements from customer i in the t^{th} time step, $t \in \{1, \ldots, m\}$. We compute the continuous voltage fluctuations as $\delta_{i,t} = v_{i,t} - v_{i,t-1}$. We also discretize these to obtain binary fluctuations, which we denote by $b_{i,t}$. Let $\mathbf{V}_i = [v_{i,t}]_{m \times 1}$ denote the m-dimensional *observation vector* that holds the time series of voltage measurements obtained from customer i. Similarly, let Δ_i and $\mathbf{B}_i$ denote the delta and binary observation vectors respectively. We assume that voltage measurements are approximately synchronised across customers, which is the case for our datasets as well.

Thus for each customer i, we have three datasets: (i) *Raw*: original data V_i, (ii) *Delta*: continuous fluctuations Δ_i, and (iii) *Binary*: discretized fluctuations $\mathbf{B}_i$. In order to identify which among the three forms of data is best suited for clustering, we consider the most common problem of partitioning customers of a feeder into 3 groups corresponding to the three phases A, B and C. We apply the K-MEANS algorithm [11] for each dataset with three measures of similarity: (i) L_2 (Euclidean) (ii) Cosine, and (iii) Correlation. To compare the resultant clusters with the true customer to phase mapping, we utilize the labelling procedure described next.

4.2 Cluster Labelling and Accuracy

To determine clustering accuracy, we assign labels to the computed clusters using an existing customer to phase mapping as follows. The computed clusters are intersected with the clusters corresponding to the existing customer to phase mapping and a one to one correspondence is established based on maximum match. Each cluster in the computed solution is assigned the phase of its matching cluster in the existing solution (Fig. 5). The accuracy of the clustering solution is then calculated as the proportion of the customers that were assigned the correct labels. Note that the same procedure may be used to calculate the clustering accuracy for experiments which determine customer to DT or feeder mapping (where the labels correspond feeders or DTs rather than phase). Additionally, the same procedure can also be used to correct an existing customer to phase or DT mapping that may be partially accurate, under the assumption that the accuracy of the existing mapping is not too low.

Fig. 6 shows the accuracy of phase clustering for the three datasets and three similarity measures for a sample feeder. We see that irrespective of the similarity measure, *Binary* performs best. Moreover, the performance of the three similarity measures with *Binary* are comparable to each other. Similar results are observed for other mapping experiments as well (customer–DT & customer–feeder). Fig. 7 shows the cluster scatter plots for two phases A and B, which compare the computed clusters with the ground truth. In each case, correlation is used as a measure of similarity. Again

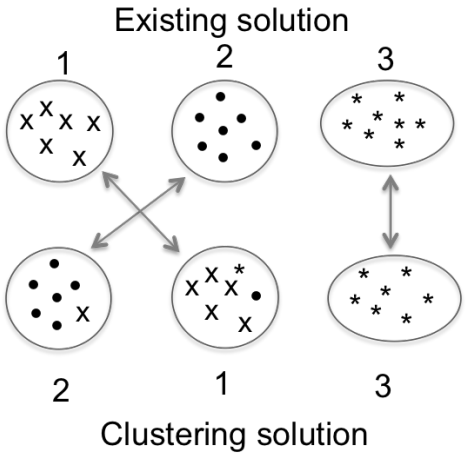


Figure 5: Labelling clusters in the computed solution using labels of an existing solution

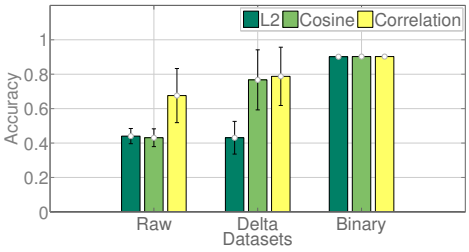


Figure 6: Clustering accuracy for the combination of datasets and similarity measures.

the plots show higher accuracy for *Binary* when compared to *Raw* or *Delta*.

In what follows in the rest of the paper, we apply K-MEANS algorithm for clustering over the binary observation vectors using correlation as a measure of similarity. Apart from K-MEANS being one of the standard techniques, it is especially suitable for our settings because, in all cases, we know the value of k beforehand. This is because we seek clustering to derive the connectivity of customers to feeder, phases or DTs and these remain unchanged despite any changes in the connectivity. For labelling the resultant clusters and calculating the clustering accuracy, we follow the same procedure as detailed above. For the experiments, we break up the datasets into batches of size 4 days and conduct clustering over the different batches, presenting the average accuracy in the results. Experiments over other batch-sizes of $1-7$ days also gave similar results.

5. CUSTOMER TO FEEDER MAPPING

This section describes our analysis towards partitioning customers belonging to different feeders. We consider three possible settings: (i) **Inter-substation:** Feeders powered by different substations, (ii) **Inter-TRF:** Feeders powered by different TRFs of same substation, and (iii) **Intra-TRF:** Feeders powered by same TRF.

For each of these settings, we combine customers belonging to the two feeders (based on the actual ground truth) and then apply voltage-data clustering (using K-MEANS) to separate them into their respective feeders. Furthermore in each case, we selectively combine all customers belonging to the same single phase (A, B or C), same two phases (AB, BC or AC) and all three phases (ABC) across the two feed-

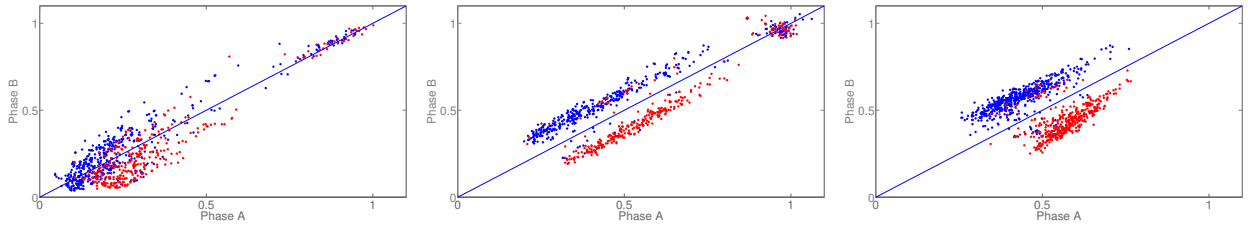


Figure 7: Cluster scatter plots. (left): *Raw* dataset, (center): *Delta* dataset and (right): *Binary* dataset

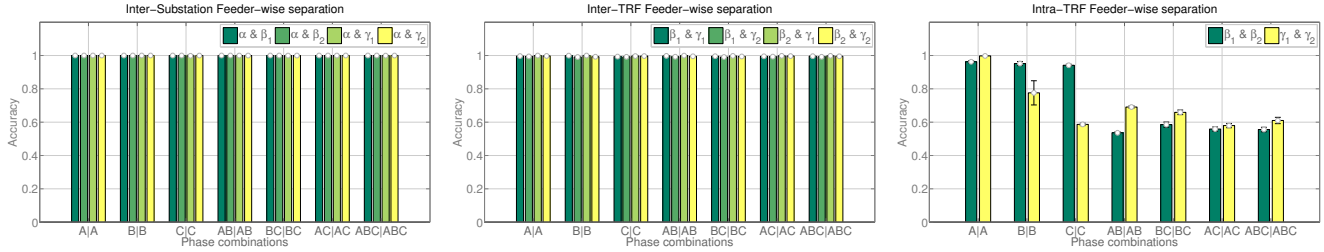


Figure 8: Feeder-wise separation of customers powered by different feeders. (left): feeders of different substations, (center): feeders of different TRFs under the same substation, and (right): feeders of the same TRF

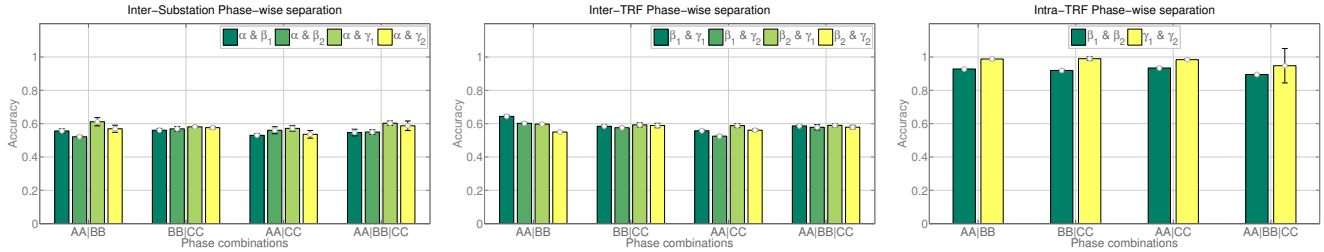


Figure 9: Phase-wise separation of customers powered by different feeders. (left): feeders of different substations, (center): feeders of different TRFs under same substation, and (right): feeders of the same TRF

ers. These experiments allow us to determine settings under which feeder correlations are stronger versus phase correlations. For example, if customers of phases A and B of two feeders (denoted by a_1, b_1, a_2, b_2) are combined and partitioned into 2 clusters, then these may separate as $\{a_1b_1, a_2b_2\}$ or $\{a_1a_2, b_1b_2\}$ (or other combinations) depending on whether customers are strongly correlated based on phase or feeders.

Feeder-wise separation. Fig. 8 shows the accuracy of feeder separation for INTER-SUBSTATION, INTER-TRF, and INTRA-TRF. The plots show average accuracy and standard deviation over multiple batches of voltage data. We see that for both the INTER-SUBSTATION and INTER-TRF setting, the accuracy of separation is almost 1, for all the phase combinations. This implies that for both these settings, voltage correlations between customers belonging to the same feeder are not only strong, but also much higher than any phase correlations. However, we see that for INTRA-TRF, where we consider feeders powered by the same TRF, the accuracy is much lower. In particular, for cases where two or all three phases are considered together, accuracy drops to 0.6. This occurs because the phase-wise correlations are quite strong among customers belonging to the same TRF, even if they belong to different feeders. Thus, the two obtained clusters do not particularly correspond to the two feeders.

The above results show that as long as customers are not powered by the same TRF, they can easily be separated out into multiple feeders using their voltage data. For instance, given customers powered by s different substations, we can partition these into s groups, one corresponding to each substation, simply by using their voltage measurements.

Phase-wise separation. For this, we conducted the same set of experiments as above, for the three settings, with the aim of obtaining clusters of customers belonging to the same phase. Fig. 9 shows the accuracy of customers being partitioned based on same phase. We observe that the accuracy is low for both INTER-SUBSTATION and INTER-TRF settings. This is expected because, as seen earlier, feeder-wise correlations are much stronger in these cases. However, for INTRA-TRF setting, the accuracy achieved is closer to 1. This goes to show that customers powered by the same TRF are related to each other much more on the basis of their phase rather than their feeder.

In summary, using voltage-data to separate customers belonging to different feeders is accurate when the feeders are powered by different TRFs. When customers belong to the same TRF, they can be accurately separated based on different phases. These results match the underlying structure of the distribution network as feeders under the same TRF are essentially powered by the same 3 phases of the TRF.

6. CUSTOMER DT MAPPING

During maintenance and repair operations, the field crew may switch customers between nearby DTs in order to restore power to the customers. Over time, such changes result in an inaccurate customer to DT mapping. In this section, we present techniques to infer customer to DT mapping solely on the basis of the data available with the utilities: GIS

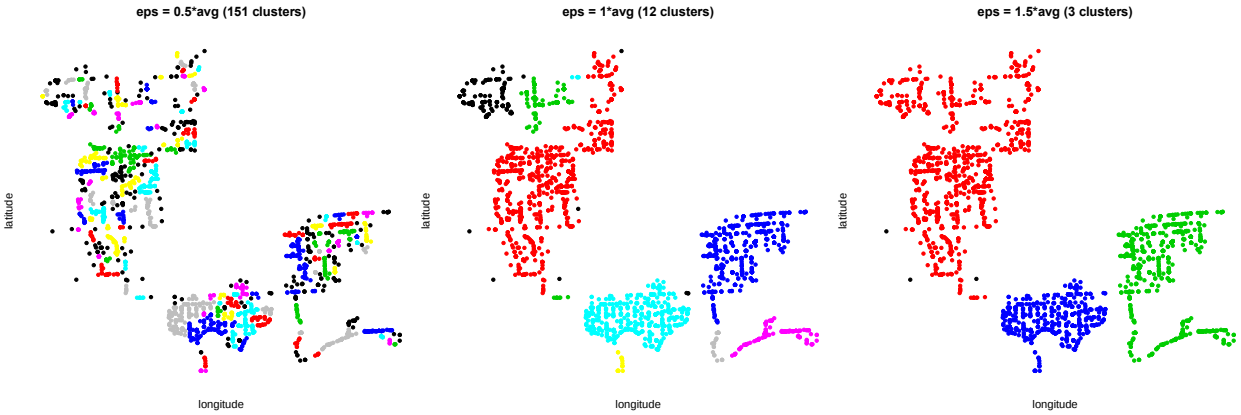


Figure 10: Anonymized spatial clusters of customers and DTs for circuit β_2

data, voltage time-series data, and any existing but partially accurate customer to DT mapping. We present results for 3 feeders: α, β_1 and β_2. Similar results are obtained for other feeders as well.

6.1 Spatial clustering for neighbourhoods

Utilities often possess GIS data of their network which includes the geospatial locations of customers and DTs. Our approach to infer customer to DT mapping leverages the GIS data to divide customers and DTs under a feeder into smaller groups within which voltage-data clustering is applied.

Now, to cluster spatial data, we require a method that is capable of discovering clusters of arbitrary shapes (corresponding to streets or blocks) without much prior information. DBSCAN [8] is a well established clustering method for such purposes. DBSCAN requires two parameters – *MinPts* and *Eps*, which determine the sizes and number of clusters respectively. A typical value of MinPts is 4 (our results do not vary with change in MinPts). On the other hand, Eps, denoting the *neighbourhood distance* of a point, plays a significant role in clustering. As our end goal is to obtain customer–DT connectivity, we used that as the factor for varying Eps. Specifically, avg denotes the average customer–DT distance in a feeder (as per the database); and Eps is varied from $0.5 * avg$ to $2 * avg$ for DBSCAN runs.

Fig. 10 shows the resulting spatial clusters (colour-coded) for one of the circuits, β_2 for three values of Eps. We see that for low value of Eps ($0.5 * avg$), a large number of clusters are obtained, most so small so as to correspond to a single street (or even less), while for higher values of Eps, we obtain well demarcated large neighbourhoods. As the purpose of clustering is to help determine customer–DT mapping, we define the accuracy of clustering as the fraction of customers that were clustered into the same cluster as their DTs. Fig. 11 (left) shows the accuracy values for the three feeders over increasing Eps values. As expected, with increasing Eps, as the cluster sizes increase, the accuracy also increases because larger clusters are more likely to contain the customers and their DTs together (tending to 1 for the full feeder scenario). Nevertheless, it is interesting to note that high accuracy (above 0.9) is achieved for Eps=avg case with only marginal improvement further on.

However, the accuracy measure by itself is not a complete indicator of the effectiveness of clustering. For the purpose of identifying customer–DT connectivity, it is preferable to

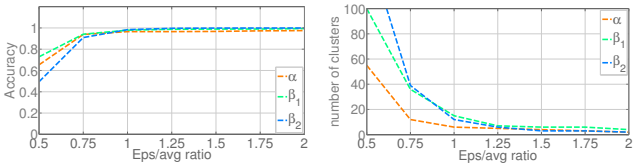


Figure 11: Spatial clustering– (left): clustering accuracy, and (right): number of clusters

have smaller spatial clusters (as long as they are accurate) so that there are fewer candidate DTs for each customer. Fig. 11(right) depicts the number of clusters generated for the three feeders for increasing Eps value. We see that while the number of clusters range from 50 to 150 (for β_2) for Eps=$0.5 * avg$, they fall rapidly to range between 10–20 for the Eps=avg case. Since the accuracy achieved by Eps=avg case is also high, it appears to be a good parameter for generating the right set of clusters.

6.2 Clustering based on Voltage-data

In order to infer the customer–DT connectivity, we apply K-MEANS clustering over the voltage data of the customers, as described in Section. 4, with number of clusters k = number of DTs. The aim is to obtain distinct clusters, each corresponding to one DT and containing only the set of those customers connected to that DT.

Given the set of customers belonging to a feeder, one may apply voltage-data clustering over them directly. However, spatial clustering of customers and DTs provides smaller clusters within each feeder. With the assumption that a customer is likely to be connected to a DT in its vicinity than to one faraway, it may be beneficial to conduct voltage-based clustering of customers within each spatial cluster separately. The final case is to consider the lateral to DT mapping. Although the lateral connectivity may not always be available, it provides good groupings of customers and DTs within each branch of the feeder. Group of customers and number of DTs belonging to a lateral are relatively small in size, however, all the customers in that group are only connected to DTs belonging to that group. Thus, lateral information provides small yet accurate spatial groups over which voltage-based clustering can be performed to obtain customer–DT connectivity.

In summary, we conducted voltage data-based clustering for all the above cases – (i) within spatial clusters (formed

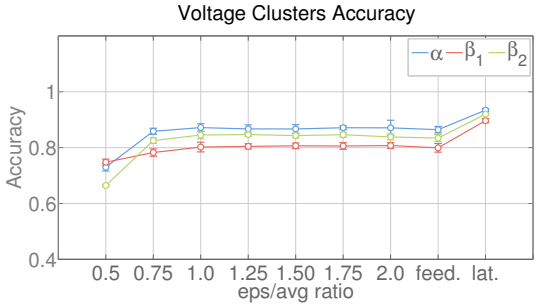


Figure 12: Customer-DT clustering: Accuracy of voltage clustering over spatial clusters formed with different Eps/avg ratios, full feeder and lateral cases.

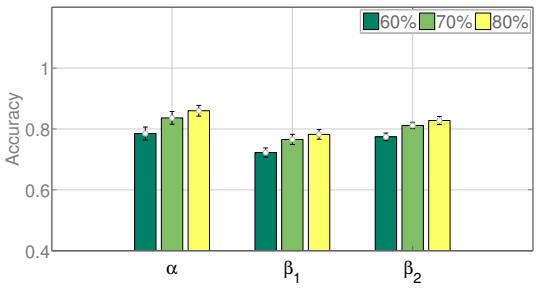


Figure 13: Customer-DT clustering: Accuracy over existing connectivity with different inaccuracies

with Eps ranging from $0.5 * avg$ to $2 * avg$), (ii) within the full feeder and, (iii) within each lateral. For each case, while using K-MEANS, k was set to the number of DTs in that group. Therefore, in the 'feeder' and 'lateral' case, it was equal to the number of DTs belonging to the feeder or the lateral respectively. However, for clustering within each spatial cluster, k value was set to the number of DTs present in that cluster. After the clusters are obtained, DTs are assigned to the clusters based on majority rule using ground truth to determine the accuracy of clustering, as explained in Section. 4.2.

As expected, when using spatial clusters generated with Eps=$0.5 * avg$ and Eps=$0.75 * avg$, the accuracy values are low, since the accuracy of the spatial clusters themselves are low (ref Fig. 11). For Eps=avg and higher, good accuracy is achieved, ranging from 0.78 for β_1 feeder to 0.87 for α feeder. For the feeder case, similar accuracy is achieved, though on average, slightly lower than when compared to the cases of Eps=avg to $1.5 * avg$. Finally, the accuracy when clustering within each lateral is even higher at around 0.9. This is expected as the lateral case represents the best-case scenario. However, lateral information may not always be available.

The previous plots show the accuracy of clustering when compared to ground truth. In practice, our goal is to correct an existing customer-DT mapping that may be partially accurate. To simulate this case, we repeat the above experiments by randomly introducing errors (20-40%) in the existing customer to DT mapping (thus leading to 80%-60% accurate ground truth scenarios). Fig. 13 plots the accuracy for full feeder case, when an existing but partially accurate mapping is used in the labeling of computed voltage-data clusters, as described in section 4.2. We observe that as long as the errors are not too high, the accuracy of the existing mapping can be improved.

In summary, the results show that by performing clustering over the voltage data and then assigning DTs to the clusters, 0.8-0.9 accuracy can be achieved depending on the setting. Moreover, we can see that when lateral-level information is not available, spatial clustering prior to voltage-clustering can lead to slightly better results on average.

7. CUSTOMER TO PHASE MAPPING

This section considers the problem of inferring customer to phase mapping, which is generally prone to more inaccuracies when compared to other mappings. As before we use voltage-data clustering to infer the phase of customers. In order to partition customers based on phases, we cluster them into three groups, *i.e.* no. of clusters $k = 3$ in K-MEANS, one corresponding to each phase A, B, and C. In this case, we compare our results with both the existing mapping available in the database as well as the true mapping that was obtained after conducting manual field inspections for the 3 feeder circuits: α, β_1 and β_2.

In order to assign a phase to each cluster, we utilize the per-phase feeder measurements obtained from the SCADA system. Let $f_{A,t}$ denote the RMS voltage measurement from phase A of the feeder during time step $t \in \{1, \ldots, m\}$ and let $\mathbf{F}_A = [f_{A,t}]_{m \times 1}$. Similarly let $\mathbf{F}_B$ and $\mathbf{F}_C$ correspond to phases B and C respectively. We pass feeder measurements $\{\mathbf{F}_A, \mathbf{F}_B, \mathbf{F}_C\}$ as centroids of the initial solution to the K-MEANS algorithm. Thus each centroid effectively has a phase. The algorithm iteratively updates the centroids and yields the final set of customer clusters. Each cluster is assigned the phase of its centroid. Note that even when data from SCADA system is not available, clustering can still be used to verify if two customers belong to the same phase. Therefore it can identify inconsistencies in the existing customer–phase mapping solely from customer voltage measurements.

As described earlier, we divide the voltage dataset, which covers 2 months, into multiple batches of 4 days each. Thus we may compute a customer to phase mapping for each batch and combine these clustering solutions to obtain the final clusters corresponding to the full dataset. We present results using two different techniques of combining these clustering solutions. (i) Majority Approach, and (ii) Cluster Ensemble.

Majority-rule approach: Each customer may be assigned a different phase based on each batch of voltage data. This simple approach assigns a final phase to a customer based on the majority of all assignments it receives over different batches.

Cluster Ensemble: There exist several cluster ensemble techniques to combine multiple partitionings of data without accessing the original feature values. Specifically, we use *Cluster-based Similarity Partitioning Algorithm* (CSPA) [13] along with k-medoid clustering algorithm to ensemble individual partitionings into a single clustering. We first generate a similarity matrix based on the individual partitionings corresponding to each batch and then apply the k-medoid algorithm to produce the final clustering. CSPA is an intuitive and efficient cluster ensemble algorithm and suits our purpose effectively.

We compare both these above methods with the accuracy averaged multiple batches in Fig. 14. We see that for all the feeders, the performance of *Cluster Ensemble* and *Majority Rule* are comparable to each other and yield close to

Table 2: Effectiveness of Phase Clustering

Feeder	Verified DTs	Positive	Negative	False Positive	False Negative	Flagged as inaccurate but incorrectly assigned	Accuracy of Clustering	Accuracy of database
α	119	100	6	5	7	1	89%	90%
β_1	200	110	78	3	8	1	94%	59%
β_2	233	161	48	13	5	6	90%	71%

Category	database solution = ground truth	database solution = computed solution	computed solution = ground truth
Positives	Yes	Yes	Yes
Negatives	No	No	Yes
False positives	No	Yes	No
False Negatives	Yes	No	No
Flagged Inaccurate but incorrectly assigned	No	No	No

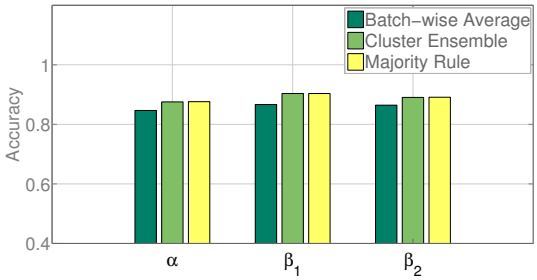


Figure 14: Customer–Phase mapping: Accuracy after applying cluster integration techniques

90% accuracy. Moreover both these methods perform significantly better than the batch-average. We obtain similar results when we use different batch sizes of $1-7$ days. Thus combining clustering solutions obtained from the different batches of voltage-data generates a final set of clusters with improved accuracy.

7.1 Comparison with Field Verified Mapping

Particularly for customer–phase mapping, we possess two mapping solutions. We possess a *Database solution* which refers to the mapping stored in the utility's database and not verified in the field. In addition, we possess the actual *Ground truth*, which was obtained after the phase of DTs was inspected in the field for the 3 feeders. Since the DTs considered are all single-phase, the customers connected to a DT have the same phase as their DTs.

We now compare our clustering-based mapping solution with manual field inspections to determine its effectiveness in identifying and correcting errors in the existing database solution. For this, we use the computed customer–phase mapping and assign a phase to each DT based on the phases of all customers under it. We then compare transformer phases across the computed solution, database solution, and the ground truth. Table 2 presents the results. The second column lists the total number of single-phase DTs which were verified in the field and the last column shows the accuracy of the existing database solution (as compared with ground truth from field inspections). The *positive* and *negative* columns count DTs wherein the computed phase is same as ground truth. The *positive* column counts DTs whose phase remains unchanged after field inspections. The *negative* column counts DTs whose phase was erroneous in the database, but was identified and corrected in the computed solution. The *false positive* counts the DTs which had an erroneous phase in the database solution but were not corrected in the computed solution. The *false negatives* counts DTs which had a correct phase in the database solution, but were identified as incorrect by the computed solution. In summary, the results show that our method of identifying customer–phase based on voltage-data clustering helped improve the accuracy of the database solution significantly in case of β_1 and β_2. For feeder α, there is a negligible drop in the accuracy. As explained above, phase is assigned to DT based on the computed phases of its customers. However, this customer–DTmapping was not verified on the field and can contain errors. These errors possibly lead to the slight drop in accuracy of the α feeder. While field inspections can be used to obtain 100% accuracy, they are expensive and time-consuming. These results show that our analytics approach is a low-cost alternative, which may be used to improve the accuracy of the database solution.

8. RELATED WORK

We know of no prior work that infers customer to feeder or transformer mapping using data already available in the distribution network. This section summarizes prior work on inferring customer phase. Caird [5] discloses a system and method for phase identification with suitably enhanced meters that can detect phases based upon a unique signal injected into the phase line. The disadvantage of signal injection methods or in general those that rely on power-line communication (PLC) is that they require enhanced hardware to transmit and receive signals at different points of the grid, increasing capital and maintenance costs. Moreover in North America, feeders from substations can run for several miles before reaching a customer. Therefore PLC-based solutions become impractical and expensive since the signal may not propagate over long distances or across DTs without repeaters. Our approach on the other hand simply relies on voltage measurements from customers and feeders and therefore does not require any additional hardware other than conventional meters.

Our prior work [1] describes an optimization approach to infer the connectivity model using a time series of customer and grid side load (kWh) measurements. The measurements are used to set up a system of linear equations based upon the principle of conservation of energy. The equations are

analyzed to fit a customer to phase mapping that is consistent with the observed time series (i.e. it minimizes the error of fit). Dilek's [7] work on phase prediction in power circuits also uses interval consumption measurements, however the author employs a heuristic Tabu search to determine the phase of attached loads. Voltage-based techniques have two advantages over load-based techniques. Firstly, voltage-based clustering techniques may be used to verify and correct an existing customer to phase mapping solely from customer meter measurements. The load-based optimization approaches require both customer and feeder measurements. Secondly, since load-based optimization techniques are based on the principle of conservation, their accuracy may reduce in the presence of any non-AMI or unmetered loads. The voltage-based approaches however are insensitive to unmetered loads in the system.

In [12], authors analyse measurements from a low voltage network consisting of a 3-phase distribution transformer and its customers. Voltage measurements of individual customers are matched with the per-phase measurements taken at the transformer to determine customer phase. Our prior work [2, 3] studies voltage measurements from a microgrid and a feeder circuit respectively to infer phase relationship. We demonstrate that individually comparing the low voltage measurements from customers with the feeder measurements yields a lower accuracy when compared to clustering, which simultaneously compares customer measurements with each other as well as with the per-phase feeder measurements. This current work, on the other hand, studies voltage data from multiple feeder circuits of a distribution network and reports accuracy results based on manual field inspections. We demonstrate the effectiveness of the clustering in comparison to field inspections for identifying and correcting phase errors. Furthermore we show that voltage measurements exhibit hierarchical correlations which can be exploited to infer customer to feeder and transformer mapping in addition to customer phase.

9. CONCLUSIONS

An accurate connectivity model is required in the planning, operations, and maintenance of distribution networks. It enables faster restoration, accurate and timely communication with customers during outages, and is also needed to efficiently integrate distributed generation sources. Automated inference of customer to distribution transformer and phase mapping using data that is already available from smart meters, feeder sensors, and GIS data allows utilities to periodically validate their connectivity and maintain a more accurate connectivity model of their distribution network without the use of expensive manual inspections.

In this work, we show that customer voltage measurements exhibit hierarchical correlations which are consistent with the hierarchical connectivity relationships that exist between customers, distribution transformers, phases, and feeders of a distribution network. We show that voltage-data clustering may be used to partition customers under different feeders as long as they are powered by different substation transformers. We present a novel analytics approach that can infer both customer-to-transformer and phase mapping with the help of customer voltage measurements in combination with feeder measurements, GIS data and any existing but partially accurate customer to transformer mapping. We showed accuracy results based on analysis of low and medium voltage measurements collected from multiple feeder circuits with more than 10K single-phase customers. Our results indicate that voltage measurements spanning several days could be utilized to infer both customer to transformer and phase mapping with high accuracy.

Future work will evaluate other clustering algorithms such as hierarchical clustering and study their performance using different cluster validity measures. We will study the use of spatio-temporal techniques which may be used to effectively combine voltage and GIS data. We will study the impact of distributed generation on customer voltage measurements and also develop approaches that combine both load (kWh) and voltage measurements to infer customer phase.

10. REFERENCES

[1] V. Arya, T. S. Jayram, S. Pal, and S. Kalyanaraman. Inferring connectivity model from meter measurements in distribution networks. In *e-Energy*, pages 173–182, 2013.

[2] V. Arya and R. Mitra. Voltage-based clustering to identify connectivity relationships in distribution networks. In *SmartGridComm*, pages 7–12, 2013.

[3] V. Arya, R. Mitra, R. Mueller, H. Storey, G. Labut, J. Esser, and B. Sullivan. Voltage analytics to infer customer phase. In *IEEE PES Innovative Smart Grid Technologies(ISGT) Europe*, 2014.

[4] J. Bouford and C. Warren. Many states of distribution. *Power and Energy Magazine, IEEE*, 5(4):24–32, 2007.

[5] K. Caird. Meter Phase Identification. US Patent Application 20100164473, January 2010. Patent No. 12/345702.

[6] G. Clark. A changing map: Four decades of service restoration at alabama power. *Power and Energy Magazine, IEEE*, 12(1):64–69, Jan 2014.

[7] M. Dilek. *Integrated Design of Electrical Distribution Systems: Phase Balancing and Phase Prediction Case Studies*. PhD thesis, Virginia Polytechnic Institute and State University, 2001.

[8] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231, 1996.

[9] J. Fan and S. Borlase. The evolution of distribution. *Power and Energy Magazine, IEEE*, 7(2):63–68, 2009.

[10] J. D. D. Glover and M. S. Sarma. *Power System Analysis and Design*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 3rd edition, 2001.

[11] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[12] H. Pezeshki and P. Wolfs. Correlation based method for phase identification in a three phase lv distribution network. In *Universities Power Engineering Conference (AUPEC), 22nd Australasian*, pages 1–7, 2012.

[13] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.