SAND2019-1988C

Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries

Logan Blakely
Sandia National Laboratories
Albuquerque, NM, 87185, USA

Matthew J. Reno
Sandia National Laboratories
Albuquerque, NM, 87185, USA

Wu-chi Feng
Portland State University
Portland, OR, 97207, USA

***Abstract* — Smart grid technologies and wide-spread installation of advanced metering infrastructure (AMI) equipment present new opportunities for the use of machine learning algorithms paired with big data to improve distribution system models. Accurate models are critical in the continuing integration of distributed energy resources (DER) into the power grid, however the low-voltage models often contain significant errors. This paper proposes a novel spectral clustering approach for validating and correcting customer electrical phase labels in existing utility models using the voltage timeseries produced by AMI equipment. Spectral clustering is used in conjunction with a sliding window ensemble to improve the accuracy and scalability of the algorithm for large datasets. The proposed algorithm is tested using real data to validate or correct over 99% of customer phase labels within the primary feeder under consideration. This is over a 94% reduction in error given the 9% of customers predicted to have incorrect phase labels.**

Keywords—machine learning, phase identification, power system simulation, spectral clustering

I. Introduction

The challenges facing utility companies as they move towards integrating distributed energy resources (DER) onto the distribution system as well as the accompanying grid modernization challenges are well known [1]. High-quality simulations are critical and rely on the current geographical information system (GIS) models of the grid, but the legacy nature of the grid and updates to the GIS system make ensuring the accuracy of those models a non-trivial task. The low-voltage side of the distribution system is particularly prone to phase labeling error [2], however, many of the smart grid research questions and modelling of residential photovoltaic (PV) systems require accurate low-voltage network models. Significant research is now going into topology estimation of the low-voltage systems in order to validate and improve the existing utility models, as discussed in [2]–[6]. Historically, improving these models has meant sending personnel into the field to do manual verification, which is often prohibitively time-consuming and expensive.

Customer electrical phase labeling is one area of distribution system topology that is known to contain errors and is critical for grid operation and moving towards higher penetrations of DER. Phase identification is important for the safety and efficiency of the grid, as well as being critical in determining placement and size of PV systems. Reference [7] discusses the importance of having balanced phases in general and [8] discusses balanced phases in the context of DER integration. The introduction of advanced metering infrastructure (AMI) smart meters, presents opportunities for a data science approach to improving and validating GIS models that did not previously exist. AMI meters provide an unprecedented amount of data on individual customer usage and time-series measurements taken at between 1-hr to 1-min intervals depending on the utility company implementation [9]. This type of large data set is ideal for the application of machine learning algorithms. Some uses of AMI data include phase identification [3], [10]–[12], load disaggregation [13], [14], and topology estimation [2]–[5]. This research proposes a spectral clustering approach, combined with a sliding window ensemble, for individual customer phase identification using only the time-series voltage profiles from the AMI data. Compared to existing literature on phase identification, the main contributions of this work are:

1) The proposed algorithm does not require substation voltage measurements. Most existing research uses the known substation phase voltages for the phase assignment. This is a major benefit of the proposed method since substation measurements (SCADA) are generally housed in a different department of the utility than AMI data. Substation voltage measurements may also be measured upstream of substation or line voltage regulators, making phase identification with voltage correlations challenging.

2) Topology information is not required for the phase identification algorithm. Some approaches use the distribution system topology information, such as customer-transformer connection labels, in the clustering algorithm itself, [10], [15]. However, this requires all other information in the distribution models to be accurate.

3) A new application of spectral clustering is presented to perform phase identification.

4) Improved accuracy and scalability of phase identification are provided using ensemble machine learning with a sliding window approach using historical data.

This spectral clustering methodology is shown to provide excellent initial results in the phase identification task, validating and improving the existing phase labels in utility models.

II. Related Work

A variety of approaches to phase identification exist in the literature. [16] uses a load summing approach, summing the individual customer loads and comparing the results to the load at the transformers and substation. This approach requires solving the linear equations produced by this method. A signal injection approach proposed in [17] requires a signal injection device as well as a device to read the injected signal. The results

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

are promising, but the addition of additional equipment adds an expense for utilizing this method.

Several approaches using different types of clustering with correlation coefficients have been attempted before. [3], [11] use hierarchical clustering and [11] introduces the sliding window approach used in this paper. [10] uses a Constrained K-Means implementation, and [15] uses a Constrained Multi-Tree algorithm to do phase identification. Both of these methods use the underlying topology as constraints in their algorithms. The customer-transformer connection labeling is used as 'must-link' constraints to reduce the number of possible pairings of customers. This approach reduces the complexity of the clustering problem but requires the assumption that all of the customer-transformer labeling is correct, otherwise this approach propagates the errors introduced by building those labels into the clustering algorithm.

Two other approaches have been recently proposed. [18] explores supervised machine learning techniques for this application. Their approach uses a field-verified subset of customers as a training dataset for the machine learning algorithm and which then predicts the remaining customers. Finally, [19] uses video imaging of light sources to group customers by phase using the oscillations in the light source due to the alternating current; this requires video imaging of individual customer buildings throughout the feeder. There have also been a variety of other applications of machine learning to power systems applications with the increase data availability [18], [20]–[22].

It is important to note that directly comparing these methodologies is difficult given the differences in datasets used for testing. Differences include synthetic data versus real-world data, varying lengths of data collection, different numbers of customers, alternative collection strategies, and differing geographic locations which introduce different seasonal conditions.

III. Data

The AMI dataset used in this research covers a 15-month period for all customers on three feeders with ~1000 customers per feeder. The data comes from the northeastern US which is notable because seasonal variance in the data may increase the difficulty of the phase identification task [3], [10]. The data comes from a utility company that has installed AMI smart meters for each customer in these feeders. The dataset contains individual AMI data for each customer. Each individual profile contains 15-minute average measurements for real power, reactive power, and voltage, as well as power generation type and a phase label (possibly incorrect) from the utility company. The power and voltage measurements are taken to an accuracy of four decimal places. The dataset for this feeder contains ~8% missing data spread throughout the customers, and this can be a challenge for certain algorithms.

In preparing the dataset before applying the clustering methodology, the voltage profiles were first normalized to a mean of 1, and then clearly erroneous values were removed from the dataset. Thresholds were set at 0.1 deviation from the mean of 1 and any value above or below that was considered an erroneous value. This may not have removed all erroneous values, but certainly any values removed in that range violate the allowable voltage range set forth in ANSI C84.1. Values adjacent to the erroneous values were also removed to reduce the chances of including inaccurate values.

IV. Methodology

The core concept behind our approach is that the voltage profiles from AMI meters can be clustered by phase using a measure of correlation or affinity between pairs of historical timeseries measurements, [3], [10], [15]. Pairs of voltage timeseries that are on the same phase will tend to have more similar variations in voltage measurements than two timeseries that come from customers on separate phases. Figure 1 shows an example of this correlation visually. There are nine examples of simplified, synthetic profiles which are grouped by correlation into three groups. Note that the key in this figure is correlation of timeseries and not simply differences in voltage magnitude.

This section is divided as follows: The Clustering Method Section details overall methodology used in this research, the Spectral Clustering section provides details on the specific clustering algorithm used here, and the Validation section describes methods used to validate these results in the absence of field verified ground truth labels.

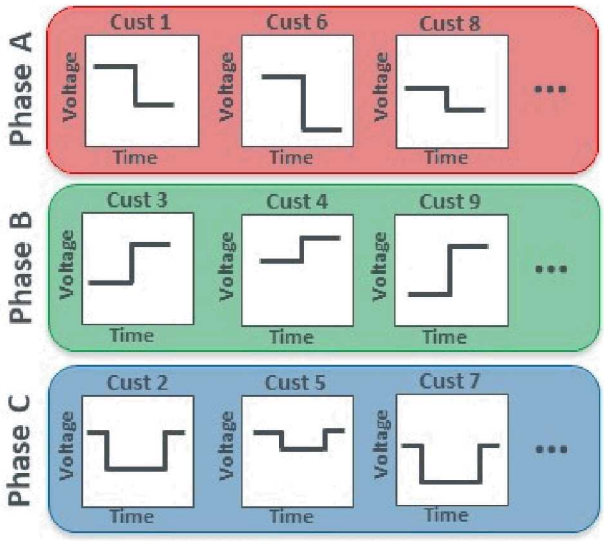

Figure 1 - Voltage correlation examples

A. Clustering Method

Our proposed machine learning algorithm for phase identification is spectral clustering with a sliding window ensemble.

Figure 2 shows a flowchart of the methodology. First, an arbitrarily sized time-window of historical data is selected. A four day window was chosen for the final implementation, based on the results in [11] and confirmed with experiments conducted in this research. Next, all voltage profiles containing missing data in that window are removed, and the remaining customers are clustered using the spectral clustering algorithm described in IV. B. The predicted phases are assigned by the majority vote of the utility labels in each of the resulting clusters. Although

the utility labels are known to have some percentage of error, the majority are assumed to be accurate. Figure 3 demonstrates this process, with the circles representing the clusters generated by the spectral clustering algorithm. Then, the subsequent window of data is selected and the process is repeated; this process repeats until all historical data is processed. For a window size of four days, this results in 121 windows for this dataset, and the final ensemble prediction is the majority vote of the predicted phases from each of the windows. In summary, there are 121 individual instances of clustering the customers without missing data during that time period. The predicted phase for each instance/window of clustering is assigned based on the majority vote of the utility labels, and the final predicted phase for each customer is assigned based on a majority vote of all the predicted phases from the 121 windows. This is shown visually in Figure 4. Each vertical box represents one window with one prediction for the phase of each customer, and then, on the right-hand side, the 'votes' from each window are used to determine the final predicted phase. The sliding window ensemble was used to handle the missing data, deal with the issue of seasonality in the data, and leverage the power of a machine learning ensemble. The ensemble both improves the algorithm accuracy, via the ensemble voting, and the scalability by not requiring the whole dataset in working memory at once.

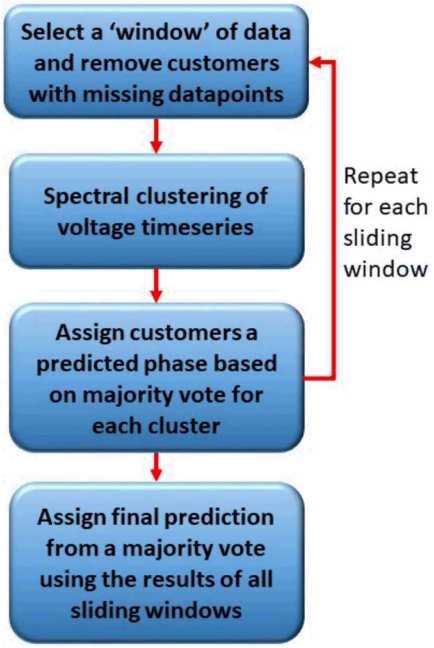

Figure 2. Clustering methodology

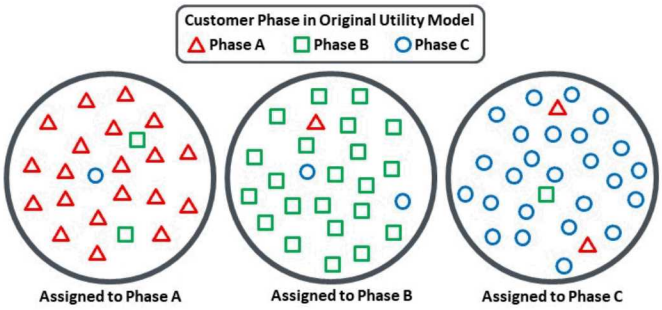

Figure 3 - Clustering prediction assignment

B. *Spectral Clustering*

Spectral clustering is a member of the unsupervised machine learning family of techniques that does not require data labels. This methodology as a whole is a hybrid machine learning approach. The utility labels are not used in the clustering itself, but they are used to assign the predicted phases to the customers after the clustering has occurred. Table 1 shows the steps in the spectral clustering algorithm. Step 2 is the primary difference between spectral clustering and other types of unsupervised learning algorithms. In the nonlinear dimensionality reduction step, the eigenvectors are computed to use as a feature representation of the data instead of using the voltage profiles directly in the clustering. For a more detailed treatment of spectral clustering please see [23].

TABLE 1- SPECTRAL CLUSTERING ALGORITHM

Spectral Clustering Algorithm	
1.	Create an affinity (similarity) matrix using a pairwise kernel
2.	Nonlinear dimensionality reduction
2.1	Compute Laplacian matrix
2.2	Compute the eigenvectors to use as feature vectors
3.	Cluster with K-Means using the eigenvectors

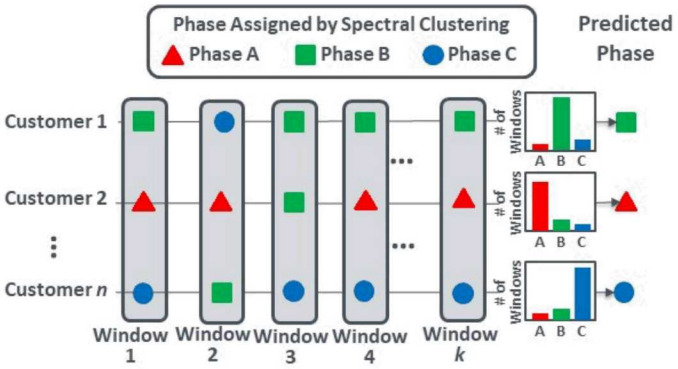

Figure 4 - Methodology

C. *Validation*

In the case of this dataset, there are no ground truth phase labels for the customer phase labels, which presents a significant challenge in validating the predicted phase results. This research uses the topology information, namely the customer-transformer connection labels as validation of the clustering phase predictions. Using the customer voltage measurements, the clustering process identifies the phase connection of the service transformers.

Figure 5 shows a Google Earth satellite image with the utility model labeling overlaid on top. There are five customers connected to the transformer, and Figure 6 shows the same transformer in Google Street view connected to the middle wire, which is Phase B by convention. In this case, the utility model phase labels and the phase labels predicted by the clustering agree that this transformer, and therefore these five customers, are connected to Phase B.

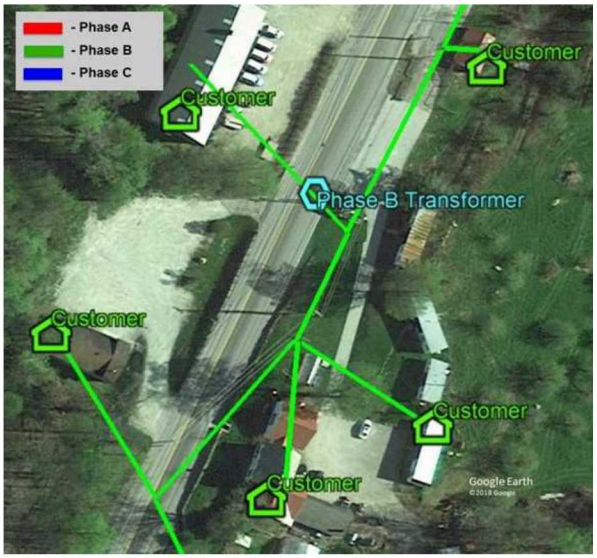

Figure 5. Google Earth view of a Phase B connected transformer serving five customers

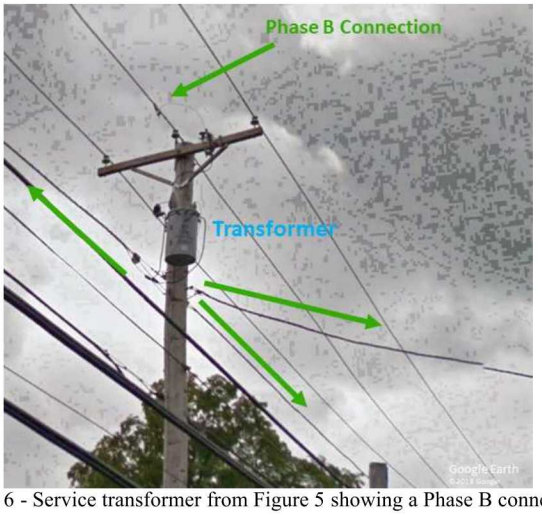

Figure 6 - Service transformer from Figure 5 showing a Phase B connection

The predicted phases are validated in a two-tier validation process using a method called 'topology validation' followed by a subset of the results validated using Google Street view. In the example above, the fact that all five of the customers seen on this transformer are predicted to be on the same phase, regardless of whether that phase matches the utility phase label or not, is considered a 'validated' prediction. Had one of the five been predicted on a different phase from the other four, that prediction would not be validated according to this topology validation metric. See the Results section and Figure 7 for further examples. It is not possible to validate all results using Google Street View due to the issues of trees, underground cables, and insufficient imagery; however, there is a subset of customers that is clearly able to be visually validated using images publicly available on Google Earth. It is important to note that without ground truth labels obtained via field verification, it is not possible to be absolutely certain of the labeling, even with this two-tier validation system. However, we believe it is possible to plausibly validate the result obtained in this research using this two-tier validation methodology.

V. Results

A. Overall Results – Feeder 1

For this feeder, ~91% of the predicted phase labels match the original utility phase labels, resulting in ~9% predicted to be on a different phase than the original utility model. Table 2 shows the results of the algorithm, with the utility phase labels on the x-axis and the clustering phase predictions on the y-axis.

TABLE 2 - RESULTS BREAKDOWN, FEEDER 1

		Utility Label			
		A	**B**	**C**	**Total**
Clustering Predicted	A	506	24	8	538
	B	10	229	3	242
	C	48	5	222	275
	Total	564	258	233	

Figure 7 shows an example of the results of the topology validation step. The lateral pictured is labeled by the utility as a Phase B lateral, as noted by the green lines. The clustering predictions and the topology validation indicate that all of the customers on all four of those transformers are predicted to all be on Phase A, instead of Phase B as the utility model has them labeled. There are fifteen customers in total represented on those four transformers, and this is a striking example of the topology validation method at work.

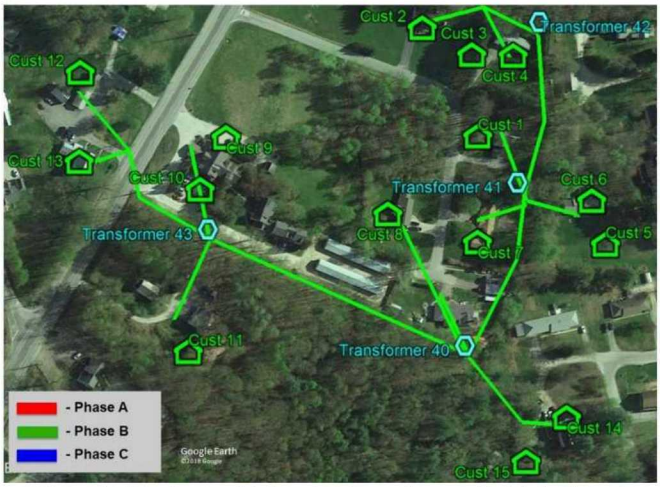

Figure 7 - Topology validation showing an incorrectly labeled lateral

The second tier of validation uses Google Street View to validate a subset of the customers that are predicted to be on a different phase from the utility phase labeling. Figure 8 shows the satellite view of the original utility labeling of Phase C (denoted by the blue line).

Figure 9 shows the Google Street View of the transformer labeled 80 in Figure 8, and here we see that the transformer is clearly connected to the middle wire, which is by convention Phase B. Thus, the predicted phase is validated by Google Street View, and there is clearly a phase labeling error in the utility model.

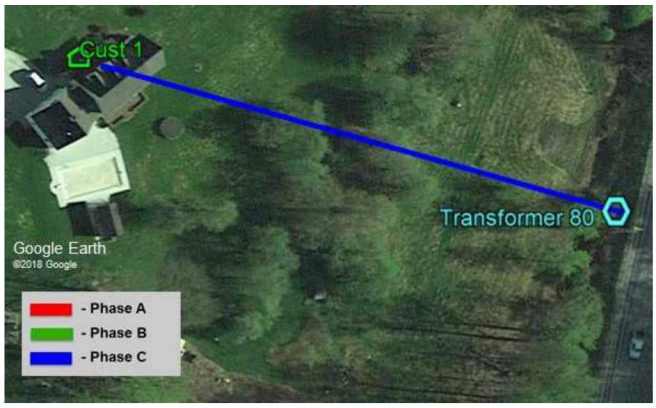

Figure 8. Single customer connected to Phase C in the original utility model (shown) and predicted to be on Phase B by the spectral clustering method.

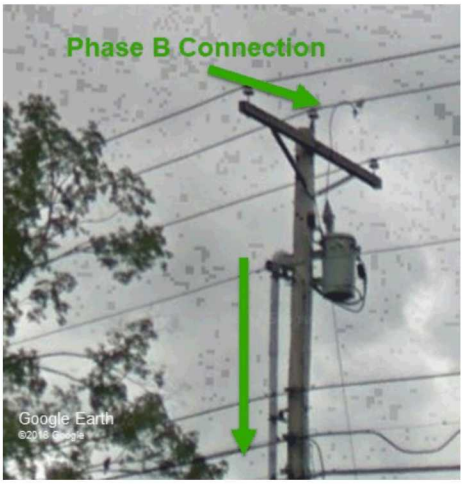

Figure 9. Transformer 80 from Figure 8 verified in Google Street View to be connected to Phase B as predicted by spectral clustering.

B. *Validated Example*

Figure 10 and Figure 11 show a second example that is fully verifiable using Google Street View images. In the utility model, shown in Figure 10, all four of the transformers on this street are labeled as being connected to Phase A. However, the phase labeling in Figure 11 is accurate based on the clustering phase predictions, the topology validation metric, and clearly visible transformer connections in Google Street view. Only transformer 51 is actually on Phase A - Figure 14, transformers 50 and 52 are on Phase C - Figure 15, and transformer 53 is on Phase B - Figure 13. Note the unlabeled home in

Figure 10, further analysis of this location revealed a mismatch between the OpenDSS model where the original data came from and the Google Earth version of the model. The OpenDSS model actually shows four more customers connected to transformer 52 than the two that are shown in Figure 10. Google Street View images demonstrate that those four customers are actually connected to transformer 53, in showing four meters visible on that unlabeled home and Figure 13 showing the Phase B interconnection for that transformer. The identification of the customer-transformer connection error is an interesting byproduct of the primary phase identification work.

Figure 10. Original utility phase labels

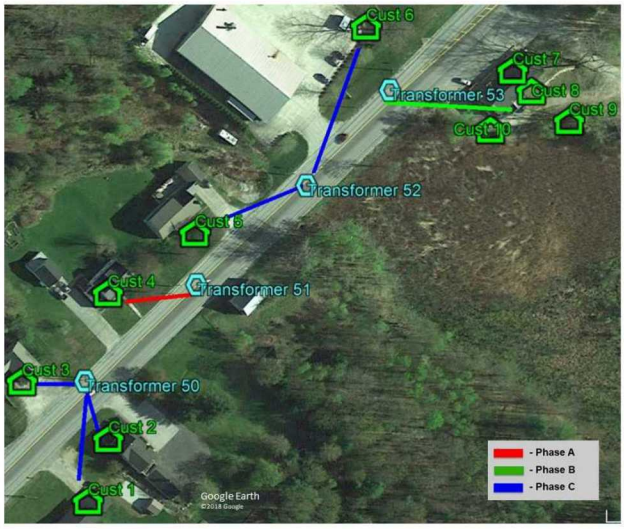

Figure 11. Spectral clustering predicted phase connections of 10 customers connected through 4 transformers as shown, validated with Google Street View.

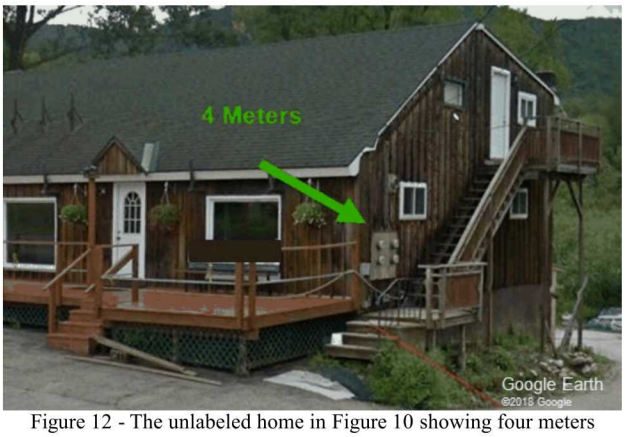

Figure 12 - The unlabeled home in Figure 10 showing four meters

Figure 13. Phase B interconnection of transformer 53

Figure 14. Phase A connection for transformer 51

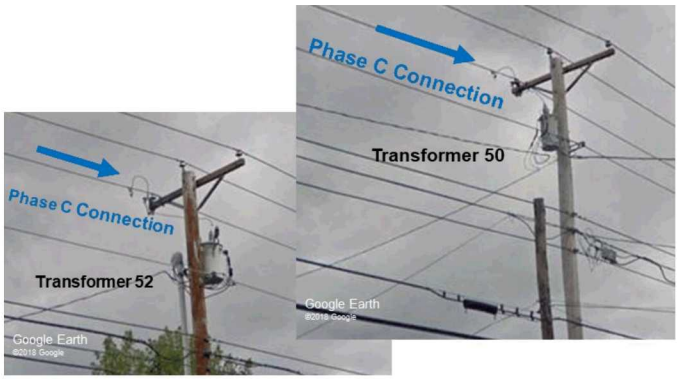

Figure 15. Phase C connections for transformers 50 and 52

C. *Summary and Comparison with Other Feeders*

Overall, spectral clustering resulted in ~91% of the original utility model customers' phase labels being correct, and ~8.5% of customers' labels that have been corrected from the utility model labeling, validated using the topology validation metric, Google Street View, and other methods. There are ~0.5% of customers where the results are unclear, and that could be a result of an algorithm prediction error or some type of other error in the utility model. Out of the total customers, ~99.5% have been plausibly accounted for using this methodology. Table 3 summarizes these results.

TABLE 3 - RESULTS OVERALL, FEEDER 1

Feeder 1 Customers	Total Customers	Validated Utility Labels	Corrected Utility Labels	Remaining Customers
Customers	1055	957	92	6
Percentages	100%	~91%	~8.5%	~0.5%

Test results also demonstrate that the spectral clustering algorithm with the sliding window ensemble is more stable than a classic k-means approach. This was shown by running a Monte Carlo simulation and comparing the individual time-window clustering instances. The baseline k-means approach failed to plausibly cluster the customers in an individual window about 6 times more often as the spectral clustering approach. The Monte Carlo simulation also demonstrated the consistency of the spectral clustering approach, showing that over 99.4% of the customers were consistently predicted on the same phase over all Monte Carlo simulations.

Table 4 and Table 5 show the results for two other feeders in the area. The results are similar overall to feeder 1, although there are more phase errors predicted in these two feeders than in feeder 1. Table 6 shows a direct comparison between the three feeders including the feeder characteristics as well as the percentage of customers that are predicted to be errors. Note that feeders 2 and 3 are more complex than feeder one, with capacitors and/or regulators. The utility noted also noted that some additional time and effort had already been expended in improving the model for feeder 1; those two factors may explain the difference in results among the three feeders.

TABLE 4 - RESULTS BREAKDOWN, FEEDER 2

		Utility Label			
		A	B	C	Total
Clustering Predicted	A	320	0	8	328
	B	33	338	1	372
	C	175	16	398	589
	Total	528	354	407	

TABLE 5 - RESULTS BREAKDOWN, FEEDER 3

		Utility Label			
		A	B	C	Total
Clustering Predicted	A	354	13	59	426
	B	26	318	1	345
	C	73	29	239	341
	Total	453	360	299	

TABLE 6 - FEEDER COMPARISON

Feeder	Voltage	Peak Load (MW)	Number of Customers	Feeder Length (KM)	Line Regulators	Capacitors
Feeder 1	12.47 kV	2.0	1153	5.4	0	0
Feeder 2	12.47 kV	1.8	1309	2.5	1 set, single-phase	450 kVAr
Feeder 3	12.47 kV	1.4	1188	3.2	0	300 kVar 300 kVar

VI. CONCLUSIONS

Spectral clustering, combined with the sliding window ensemble approach correctly identified the phase connections of customers, validating and improving the existing utility model for this section of the distribution system. It was shown on feeder 1 that the utility phase tracking was fairly accurate with ~91% of the phase labels in the original utility model being validated with voltage AMI measurements. ~8.5% of phase labels were corrected. This is a ~94% reduction in the total ~9% error predicted to be in the original utility model phase labels and a ~99.5% reduction in the total uncertainty of the model. Uncertainty in existing utility models for the low-voltage portion of the distribution system is a major limiting factor in grid modernization and the continuing addition of DER technologies onto the power grid. A novel spectral clustering algorithm was demonstrated on three distribution feeders as a promising technique to remove uncertainty in the phase labeling of customers, as well as demonstrating potential uses of machine learning in leveraging the big data that is being produced by AMI meters.

REFERENCES

[1] B. Palmintier *et al.*, "On the Path to SunShot: Emerging Issues and Challenges in Integrating Solar with the Distribution System," *Natl. Renew. Energy Lab.*, vol. NREL/TP-5D00-65331, 2016.

[2] W. Luan, J. Peng, M. Maras, B. Harapnuk, and J. Lo, "Smart Meter Data Analytics for Distribution Network Connectivity Verification," *IEEE Trans. Smart Grid*, vol. 6, p. 1, Jul. 2015.

[3] T. A. Short, "Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 651–658, Jun. 2013.

[4] J. Peppanen, S. Grijalva, M. J. Reno, and R. J. Broderick, "Distribution System Low-Voltage Circuit Topology Estimation using Smart Metering Data," *IEEE PES Transm. Distrib. Conf. Expo. Dallas TX*, 2016.

[5] X. Zhang and S. Grijalva, "A Data-Driven Approach for Detection and Estimation of Residential PV Installations," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sep. 2016.

[6] M. Lave, M. J. Reno, R. J. Broderick, and J. Peppanen, "Full-Scale Demonstartion of Distribution System Parameter Estimation to Improve Low-Voltage Circuit Models," presented at the IEEE Photovoltaic Specialists Conference (PVSC), 2017.

[7] J. Zhu, M.-Y. Chow, and F. Zhang, "Phase Balancing using Mixed-Integer Programming," *IEEE Transcations Power Syst.*, vol. 13, no. 4, Nov. 1998.

[8] R. Yan and T. K. Saha, "Voltage Variation Sensitivity Analysis for Unbalanced Distribution Networks Due to Photovoltaic Power Fluctuations," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 1078–1089, May 2012.

[9] "Advanced Metering Infrastructure and Customer Systems: Results From the Smart Grid Investment Grant Program," *US Dep. Energy Off. Electr. Deliv. Energy Reliab.*, Sep. 2016.

[10] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 259–265.

[11] R. Mitra *et al.*, "Voltage Correlations in Smart Meter Data," *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1999–2008, 2015.

[12] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2012, pp. 1–7.

[13] G. Zhang, G. G. Wang, H. Farhangi, and A. Palizban, "Data Mining of Smart Meters for Load Category Based Disaggregation of Residential Power Consumption," *Sustain. Energy Grids Netw.*, vol. 10, pp. 92–103, Jun. 2017.

[14] X. Jin and D. Christensen, "Disaggregating Smart Meter Data to Identify Electric Loads and Control Opportunities," *Non-Intrusive Load Monit. NILM*, 2018.

[15] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, "Phase Identification of Smart Meters by Clustering Voltage Measurements," *Power Syst. Comput. Conf. PSCC*, 2018.

[16] V. Arya *et al.*, "Phase identification in smart grids," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2011, pp. 25–30.

[17] K. J. Caird, "Meter Phase Identification," *US Pat. 8143879*, Mar. 2012.

[18] B. Foggo and N. Yu, "A Comprehensive Evaluation of Supervised Machine Learning for the Phase Identification Problem," *World Acad. Sci. Eng. Technol. Int. J. Comput. Syst. Eng.*, vol. 12, no. 6, 2018.

[19] M. Sheinin, Y. Y. Schechner, and K. N. Kutulakos, "Computational Imaging on the Electric Grid," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2363–2372.

[20] E. M. Stewart *et al.*, "Integrated Multi-Scale Data Analytics and Machine Learning for the Distribution Grid and Building-to-Grid Interface," *Lawrence Livermore Natl. Lab.*, vol. LLNL-TR-727125, 2017.

[21] L. Blakely, M. J. Reno, and R. J. Broderick, "Decision Tree Ensemble Machine Learning for Rapid QSTS Simulations," *IEEE Innov. Smart Grid Technol. ISGT*, 2018.

[22] M. J. Reno, R. Broderick, and L. Blakely, "Machine Learning for Rapid QSTS Simulations using Neural Networks," presented at the IEEE Photovoltaic Specialists Conference (PVSC), 2017.

[23] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Proc. 14th Int. Conf. Neural Inf. Process. Syst. Nat. Synth.*, pp. 849–856, Dec. 2001.