Abstract: Calibrating distribution system models to aid in the accuracy of simulations such as hosting capacity analysis is increasingly important in the pursuit of the goal of integrating more distributed energy resources. The recent availability of smart meter data is enabling the use of machine learning tools to automatically achieve model calibration tasks. This research focuses on applying machine learning to the phase identification task, using a co-association matrix-based, ensemble spectral clustering approach. The proposed method leverages voltage time series from smart meters and does not require existing or accurate phase labels. This work demonstrates the success of the proposed method on both synthetic and real data, surpassing the accuracy of other phase identification research.

1 Introduction

The quality of distribution system models is increasingly important as the need for accurate and granular simulations escalates. The potential for high penetrations of distributed energy resources (DER) to negatively affect the stability of distribution systems rises as the percentage of DER increases. Highly-accurate simulations are a critical factor in mitigating such risks [1–3]. Rapid and accurate hosting capacity analysis is required for integrating large percentages of DER into the distribution system, requiring accurate models of those systems. For hosting capacity analysis, computation time and precision level are becoming bottlenecks for the installation of photovoltaic (PV) systems and other DER [4–6].

Unfortunately, the distribution system, particularly the low-voltage side, has historically been prone to larger errors than the transmission side of the grid [7]. Distribution system errors include errors within the medium-voltage portion such as topology errors [8–11]; and errors within the low-voltage portion such as phase label errors [12–14]; errors in model parameters (wire type, overhead line configuration etc.) [9, 12, 15, 16]; and behind-the-meter parameters, such as for residential PV installations [17–19]. Blakely et al. [20] provide a detailed overview of commonly found errors in distribution systems with examples and a literature review of current work being done in these areas.

The advent and popularity of advanced metering infrastructure (AMI) have made available a range of possibilities for data analysis that were previously unavailable. The increase in the amount of data enables the application of data science and machine learning approaches for model calibration without the expense of manual verification. The authors of [21, 22] discuss AMI and smart metering deployment, and [16, 23, 24] discuss some of the specific data requirements for model calibration tasks and the array of data considerations for AMI data collection and analysis in general.

This work focuses on the task of customer phase identification using AMI data. Accurate phase labelling is one aspect of ensuring accurate hosting capacity analysis simulations. Distribution system models often have erroneous or missing phase labelling – if the phase information was recorded at all. This work aims to calibrate the distribution system model phase labels using a data-driven, machine learning approach without the need for manual verification by field personnel. This approach has been tested on both synthetic and real feeder data. Contributions of this work include:

(i) Phase labelling is not required for this algorithm to function. This algorithm works without requiring any initial accurate labelling of the customer phases, although if labelling exists it can be optionally used in the final step.

(ii) Data streams other than AMI voltage time series are not required for the phase grouping. This method does not require supervisory control and data acquisition system (SCADA) measurements, topology information, or power measurements.

(iii) This method improves upon the accuracy of similar existing methods, [14, 25].

2 Related works

The phase identification task is well documented in the literature, and there are a wide variety of approaches. There is not a consensus on the best approach or algorithm, and it is difficult to directly compare the algorithms due to the differing datasets used in the research. Three general categories of approaches are those that use additional hardware, those that primarily using power measurements, and those primarily using voltage measurements.

Of the three, the last has had the most research attention. The research in [26] is an approach that requires additional hardware. This is a signal injection methodology where distinct signals are injected into each of the phases and then, by reading that signal at the customer level, the phases are assigned. The authors of [27, 28] utilise two different variations of microsynchrophasor technology to take readings at locations of interest in the field. These methods are extremely accurate, but they are both manual methods that require extensive work by field personnel. Sheinin et al. [29] describe a video-based approach. In this methodology, video footage of lighting from buildings is taken and then analysed for characteristics undetectable to the human eye. Those characteristics are then used to group buildings by their phase. The primary disadvantage of the hardware-based methods is the added expense and time associated with the hardware devices themselves.

One advantage of using power-based methods is that, currently, power data is more readily available than voltage data. Arya et al. [30] use the power measurements from customers for the phase identification task. This approach is based on the fact that less line loss and unmetered load, the sum of all loads on one phase should be equal to the load on the same phase measured at the substation. One disadvantage to this type of power-based methods compared to voltage-based methods is their sensitivity to unmetered loads and scalability to the number of customers. The study by Pappu et
al. [31] is based on a similar approach that uses principal component analysis and a graph theory methodology. Xu et al. [32] use a method based on spectral saliency analysis on load profiles. Load profiles are converted to the frequency domain and filtered, and then "salient" events, large changes, are identified and matched in time with the load profiles at a substation. Tang and Milanovic [33] use the active power measurements from smart meters and the identification with a least absolute shrinkage and selection operator (LASSO) method for the phase identification task. These methods also have the disadvantage of requiring smart meter power measurements or substation data.

There are a significant number of instances in the literature using the voltage AMI time series, most of which, including the method proposed here, leverage the correlation between pairs of voltage profiles of customers on the same phase. Customers who are connected to the same phase will have voltage measurements that are more correlated to each other than two customers which are connected to different phases. Supervised machine learning is investigated in [34, 35]. Foggo and Yu [36] use a subset of manually verified customers as training data for the supervised machine learning and then use the machine learning to predict the phase of the other customers on the feeder. Choosing which customers to manually verify and how many are required is part of the research presented there. Foggo and Yu [36] improve on that approach by using information losses and entropy to guide the machine learning algorithm. Both of these methods still incur the expense of field verification of a subset of customers in each feeder.

Short [12] uses an approach based on pairwise linear regression. This approach has also been applied to the parameter estimation problem [15, 16]. This method shares the same disadvantage as the power methods in that it is sensitive to unmetered loads.

Pezeshki and Wolfs [36] use a cross-correlation between customer voltage profiles and known voltage profiles for each phase. This approach requires having known voltage profiles, perhaps from the substation, for each phase.

Hierarchical clustering was applied to this task using correlation coefficients in [12, 13]. Mitra et al. [13] also provide some key insights that are used in the proposed algorithm including the voltage difference representation and the window ensemble approach. Initial testing using correlation coefficients in this way was investigated produced unsatisfactory results with low accuracy on our datasets. Another method using correlation coefficients is [37]. This work focuses on parameter estimation, building up a topology tree using point-of-coupling calculations and correlation coefficients. After building the topology tree, they propose using the tree for the phase identification task. This work requires both real and reactive power for the point-of-coupling calculations.

Some approaches include topology constraints in their approach. Olivier et al. [38] use a constrained multi-tree approach that incorporates known topology information as constraints to the phase identification. Wang et al. [39] incorporate the transformer labelling as must-link constraints in a constrained K-means algorithm. The demonstrated results are excellent, although they do rely on the accuracy of the included topology constraints. Otherwise, the errors there are propagated through into the phase identification results. Wang and Yu [40] formulate the phase identification problem as a maximum marginal likelihood estimation problem that utilises both smart meter data (voltage and real and reactive power) as well as topology information from the feeder. This method relies on accurate knowledge of the feeder topology and three smart meter data streams.

Ng et al. [41] provide an in-depth description of the spectral clustering algorithm and clustering ensemble which have a significant role in the phase identification method. Second, the phase identification algorithm itself is described in detail.

3.1 Spectral clustering process

Clustering, as a group of methodologies, is considered an unsupervised machine learning technique. This means that labelled data points are not required to use the method. K-means and hierarchical clustering are two other examples of unsupervised machine learning methodologies. The proposed methodology uses only voltage time series measurements from AMI meters, i.e. a single stream of data. That data is transformed in two ways. First, the voltages are converted into a per-unit representation using the ideal mean for the time series. Secondly, the time series are converted into a ‘voltage difference’ representation by taking the difference between adjacent measurements of the time series. This results in a time series where the values represent the change in voltage at each time step. The remainder of this section is structured as follows. First, there are brief descriptions of the spectral clustering algorithm and clustering ensemble which have a significant role in the phase identification method. Second, the phase identification algorithm itself is described in detail.

![Fig. 1 Methodology flowchart](Image 349x83 to 513x291)
implementation used in this research is the Python Scikit Learn implementation [42].

In general, spectral clustering calculates (or accepts as input) a pairwise affinity matrix between samples, computes the eigenvectors, and then clusters the data into a user-defined number of clusters using the eigenvectors. A Laplacian matrix is computed from the affinity matrix; this matrix will be approximately block-diagonal and from there the eigenvectors are computed. A subset of the eigenvectors is then used for the clustering step. In the Scikit Learn implementation there are two choices for the clustering step, either 'k-means' or 'discretise'. Spectral clustering provides a non-linear dimensionality reduction of the input data which differentiates it from other conventional clustering algorithms.

The proposed phase identification algorithm uses slightly different versions of spectral clustering in two locations. As shown in Fig. 1, spectral clustering is used in Step 2. In that instance of the spectral clustering algorithm, the voltage-time series data is used as input and the spectral clustering algorithm uses a radial basis function kernel to calculate the pairwise affinity matrix. The $k$-means spectral clustering option is used for the clustering step. More details on $k$-means can be found in [43]. Please see Section 3.3 for a discussion of the number of clusters.

In Fig. 1 Step 4, spectral clustering is used again. The requirements for clustering in this step are different from the requirements in Step 2, thus a different parameterisation of spectral clustering is used. The primary differences are that there is a precomputed affinity matrix available and a different number of clusters is used. The co-association matrix generated by previous steps in the phase identification algorithm (see Section 3.2) is used as input as a ‘precomputed affinity matrix’ and the spectral clustering algorithm proceeds directly to calculating the Laplacian matrix and the eigenvectors. The ‘discretise’ option is used for the clustering step. This option produced better results during testing than the $k$-means version of spectral clustering. $K$-means is known to be sensitive to initialisation and we believe this to be the key factor in the ‘discretise’ option achieving better performance in this final clustering step. This method turns the clustering into an optimal discretisation problem; more details can be found in [44].

Spectral clustering is applied in this paper using a cluster ensemble technique. A cluster ensemble is the aggregation of a number of distinct clustering instances for use in a final ‘consensus clustering’ algorithm. In particular, this work uses a co-association matrix method for combining the results of individual clustering instances. For an overview of cluster ensemble techniques please see [45]. The work in [46] demonstrates a cluster ensemble methodology using spectral clustering in the image segmentation domain. That work leverages similar principles to the work demonstrated here; however, the algorithm and implementation differ significantly. The cluster ensemble technique corresponds to Steps 3 and 4 in the following section.

3.2 Phase identification algorithm

This section describes the phase identification algorithm in detail, and the numbered sections follow the steps shown in Fig. 1 flowchart.

(i) First, the available historical data (voltage time series for each customer) are divided into 4-day periods or ‘windows’ and all customers with missing data during that period are removed from that window. This process accomplishes several goals. It allows the use of all available historical data, only requiring a subset of the data for computation at one time. The window approach gives a method for dealing with missing data within the dataset. The ensemble nature of the window approach leverages the power of ensemble methodologies and helps account for variations in the dataset due to seasonal variability, customer load variability, or other factors. This step was based on research in [13, 14]. In each of steps 2 and 3 of the algorithm, only a single window is considered. See Section 3.3 for a more detailed discussion of the window size parameter.

(ii) Secondly, the remaining customers within that window, the ones with complete data during that window, are clustered using the spectral clustering algorithm. Six, twelve, fifteen, and thirty clusters were chosen for use during this step. This step returns a set of cluster labels from each clustering instance. The results from four clustering instances are used for Step 3. See Section 3.3 for a discussion of the number of clusters used in this step. Foundational research for this step can be found in [14, 25]. However, both of those works use only a single value for the number of clusters.

(iii) The cluster label information from Step 2 is then used to populate a co-association matrix. A symmetric $n \times n$ matrix $A$ was created and initialised to zeros as a pre-processing step, where $n$ is the total number of customers and $A_{ij} = A_{ji}$ is the weight between customers $i$ and $j$. This weight represents the affinity between those customers; customers with larger affinities will have been clustered together in Step 2 more often than customers with low affinities. The matrix is updated with the spectral clustering results from each window. Each cluster produced by the spectral clustering algorithm is represented as an adjacency matrix which is used to increment the corresponding field in the co-association matrix. This step is shown in Fig. 2 and the update step is shown in Fig. 3. For example, if customers 1 and 2 were clustered together, then the field in the matrix, row 1 and column 2 (as well as row 2 column 1) would be incremented to a larger weight. In practice, each update is done in two cells of the co-association matrix due to the symmetry of the matrix. Thus, in each window, customers which are clustered together are updated to have stronger weights in the co-association matrix. At this stage the algorithm repeats steps 1–3 for the subsequent window of data, as shown in Fig. 1, and the windows on the left side of Fig. 2. After all, data has been processed in this way, meaning all available windows have been used, the result is a co-association matrix that contains the pairwise clustering information from all of the instances of spectral clustering. An alternative way of viewing this matrix is as a histogram which counts the number of instances where a customer $i$ was clustered with a customer $j$. This step represents a novel contribution of this algorithm.

(iv) Once all available data have been used, the co-association matrix is normalised to account for the influence of missing data on the matrix. This is necessary because each cell in the co-association matrix may have received a different number of increments due to customers being removed in windows due to missing data. A count is maintained through all available windows, counting the number of windows where each pair of customers was present in the window, meaning those customers could have been clustered together. This count is accumulated in a symmetric matrix of the same size as the co-association matrix. Each cell in the co-association matrix is divided by the corresponding cell in the count matrix to form a normalised version of the co-association matrix; this is shown on the right side of Fig. 3. The final step uses the normalised co-association matrix as input to another spectral clustering algorithm for the partitioning of the customers into final clusters representing the phases; this is shown in Fig. 4. After the final clusters have been obtained, there are two choices in terms of finishing the phase identification task. If utility labels exist, and the majority are deemed accurate, the final clusters could be labelled using a majority vote of the utility labels contained within those clusters. However, this does require most of the labels to be accurate. Another possibility is to end the method at the determination of those clusters and leave the final determination of phase labels for the utility. This could either be done by comparing to known via manual verification. The actual final number of clusters used in this step may depend on the characteristics of the feeder in question. Please see Section 3.3 for a more detailed discussion of the number of final clusters. Step 4 also represents a novel contribution of this algorithm. The use of the cluster ensemble with the co-association matrix is one of the key differentiators between the proposed method and the work in [14]. This significantly increases the robustness of the algorithm and removes the reliance on existing utility phase labels as shown in the Results Section 4.
This section discusses the selection of the number of clusters, both in stage 2 and stage 4 shown in Fig. 1, and the selection of the window size in stage 1 of Fig. 1.

The number of clusters for Step 2 of the algorithm was obtained through testing of a variety of values. The number of clusters needs to be large enough to allow variation in the clustering and the variability in subsequent windows is desirable to populate the co-association matrix. These considerations are balanced by the fact that the number of clusters must be small enough to adequately populate the co-association matrix in Step 3 using the available data. Using a high number of clusters would require having access to more data due to the tendency of some isolated clusters to form between highly correlated customers. Ideally, all customers on a particular phase will have similar weights and having a smaller number of clusters facilitates this. Balancing these considerations dictates the range of possible values for the number of clusters. Fig. 5 shows results on the synthetic dataset for various numbers of clusters. One month of the synthetic data with large levels of measurement noise was used to illustrate this.

**Fig. 2** Create adjacency matrix based on the results of a window clustering using 5 clusters

**Fig. 3** Updating the co-association matrix with the results of each individual window clustering and then normalising the co-association matrix by dividing by the matrix tracking the customer presence in each window

**Fig. 4** Final cluster assignment of which customers are on the same phase as each other. This is followed by the optional step of using the utility phase labels for the final phase assignment

**Fig. 5** Phase identification accuracy for the synthetic data using the ensemble clustering with different numbers of clusters for Step 2 of the algorithm

3.3 Parameter selection discussion

This section discusses the selection of the number of clusters, both in stage 2 and stage 4 shown in Fig. 1, and the selection of the window size in stage 1 of Fig. 1.

The number of clusters for Step 2 of the algorithm was obtained through testing of a variety of values. The number of clusters needs to be large enough to allow variation in the clustering and the variability in subsequent windows is desirable to populate the co-association matrix. These considerations are balanced by the fact that the number of clusters must be small enough to adequately populate the co-association matrix in Step 3 using the available data. Using a high number of clusters would require having access to more data due to the tendency of some isolated clusters to form between highly correlated customers. Ideally, all customers on a particular phase will have similar weights and having a smaller number of clusters facilitates this. Balancing these considerations dictates the range of possible values for the number of clusters. Fig. 5 shows results on the synthetic dataset for various numbers of clusters. One month of the synthetic data with large levels of measurement noise was used to illustrate
the effects of the cluster size parameter choice. The leftmost column shows the results for a single window (i.e. without the ensemble). From this set of results, it is clear that using multiple clusters is advantageous in the ensemble. Although the smaller numbers of clusters show decent results, they are not perfect. Furthermore, choosing a small number of clusters in this step forces a choice for the final number of clusters. The next paragraph discusses the challenges of choosing a specific number of final clusters. Utilising multiple numbers of clusters in this step leverages the advantages of small numbers of clusters as well as a large number of clusters. Diversity in the clustering instances that are used in the cluster ensemble is critical for the success of the method. [45, 46], and altering the parameters of the initial clustering instances is one way to achieve that diversity. Further testing also showed that using multiple values for the number of clusters in each window also resulted in a more consistent co-association matrix between isolated runs of the algorithm.

The number of clusters used for the final clustering in Step 4 is driven by the number of expected phases in the feeder. This value will change depending on the characteristics of the feeder. For example, in the synthetic data results shown in Section 4, it is known in advance that there are three phases on that feeder, i.e. there are no delta connected transformers. However, if there were delta connected transformers a different final cluster value would be required. Although not directly investigated in this work the presence of voltage regulators may require additional consideration in the choice of the final number of clusters; however, that is left to future work. Similarly, the utility labels from the dataset shown in Section 5 contain the labels ‘Phase A’, ‘Phase B’, ‘Phase C’, and ‘Unknown’. In actuality, some of the customers labelled ‘Unknown’ are likely on a different feeder altogether. Thus, the final number of clusters needs to be 4 in this case. Based on the known, or estimated, characteristics of the feeder, the number of final clusters should be set dynamically.

The window size in this work was set at 4 days. The considerations on this choice balance two competing factors. In our experiments, larger window sizes tended to achieve better performance. However, a critical aspect of this algorithm is the ability to deal with missing data by removing customers with missing data during a particular window. Using a 4-day window results in 91 windows over a one-year period, whereas a 7-day window results in 52 windows, and a one-month period results in just 12 windows. As the window size grows larger, more and more customers are lost in each window due to missing data. Fig. 6 shows the results of a set of simulations varying the window size. In each simulation, 22 windows were used. The effects of the window size can be seen in the accuracy on the y-axis. Smaller window sizes show slightly decreased accuracy. The large decrease in accuracy after the 7-day mark is due to customers being completely lost due to missing data. Thus, 4-days was chosen to balance these factors. 4-day windows were also used successfully in [14]. In [13], 4-days was chosen for the final version of the algorithm, and the authors note that choosing window sizes between 1 days and 7 days resulted in the similar performance of their algorithm.

Fig. 6 Accuracy of different window size in the cluster ensemble using only 22 windows to establish the co-association matrix

The tuning of these algorithm parameters was accomplished by leveraging the randomness in the data manipulations discussed in Section 4.1. The data manipulations, such as measurement noise, meter bias etc. were used to instantiate new versions of the dataset for each test. In addition, the results in Section 5 demonstrate the final parameters on a second, utility, dataset.

4 Synthetic dataset results

This section contains the results for experiments conducted on a synthetic dataset where the ground truth labels are known. This allows for rigorous testing and manipulation of the dataset to create a more realistic test environment.

4.1 Dataset

The synthetic dataset used for these simulations is described in detail in [16] and is also used in the AMI data analysis in [23]. The dataset consists of 12 months of AMI data for 1369 customers simulated on the Electric Power Research Institute’s (EPRI) Ckt5, Fig. 7. [47] The synthetic dataset allowed simulation of baseline results with ground truth phase labels for all customers and voltage time series free of measurement noise or other data issues. Realistic data issues were later added to the dataset for testing purposes.

Two types of data manipulations were used to further test the algorithm and simulate more realistic data conditions. The first type of data manipulation concerns AMI data collection methods such as the measurement interval, and meter resolution. Other data manipulations were related to possible data issues such as meter bias, measurement noise, and missing data. A combination of these factors was combined to create a test case by injecting uniformly distributed measurement noise up to 0.02% of the mean, injecting uniformly distributed meter bias of up to 0.02% of the mean, and by removing 0.2% of the measurements to simulate missing data. The AMI data collection methods used were 15-min average measurement interval, 0.1 V resolution measurements, 6 months of available data, and AMI meter penetration of 100%. Further details of the specific implementation of the data manipulations used can be found in [23]. The values used in the results shown in Figs. 8–10 reflect the ‘Test Case’ values discussed there. Furthermore, a set of customers was chosen to intentionally force their phase labels to be incorrect to simulate model errors. Various levels of mislabelling were chosen to test the effects of the quantity of mislabelled customers on algorithm performance.

4.2 Results

Fig. 8 shows confusion matrices corresponding to testing the algorithm with 25% of the phases mislabelled. The top half of the figure shows the distribution of phases that were intentionally mislabelled for this simulation, and the bottom half of the figure...
Table 1 shows the results of testing to determine the robustness of the proposed algorithm. Measurement noise is simulated as uniformly distributed noise up to a maximum error in the real power measurements. These results use the dataset characteristics described in Section 4.1, testing varying levels of measurement noise. The values shown are the averaged results of five independent simulations. We can see that the algorithm is relatively robust to measurement noise in the dataset. There are no errors in phase predictions until 0.65% maximum noise and, at that level of noise, there are not consistent errors. That is driven by the randomness associated with the location and quantity of the noise within each simulation.

### 4.3 Results comparison

The proposed algorithm is compared to a phase identification algorithm with a similar approach. For algorithm, details see [14]. The same experiment shown in Table 1 was conducted for the comparison algorithm in [23]. Those results suggest that the measurement noise should be <0.25% (which approximately corresponds to the 0.2-meter class). The results shown in Table 1 demonstrate considerable improvement, up to ~0.65% maximum noise before errors begin to occur.

Figs. 9 and 10 show the results of the testing on the synthetic dataset for both the algorithm in [14], listed as ‘Comparison Method’ as well as the proposed algorithm. In both figures, the results shown are from a Monte Carlo simulation consisting of 500 individual simulations, where the only difference between each simulation is which customers were chosen to be mislabelled. The minimum, maximum, and mean percentages from the 500 simulations are shown for the comparison algorithm. For the proposed algorithm, the results for all three metrics (min, max, mean) are identical so only one line is plotted. The x-axis of Fig. 9 shows the percentage of customers that have been injected with phase label errors for the purposes of testing, and the y-axis shows the percentage of those mislabelled customers that have been corrected. For example, for the minimum accuracy of the comparison method at 35% of customers mislabelled the method was able to correctly identify ~90% of customer phases; 35% of 1369 customers is 479 customers. Therefore, 48 of those 479 customers remained incorrect after the algorithm completed. In contrast, the proposed algorithm retains 100% accuracy for all 500 simulations. Note the unevenness of the line corresponding to the minimum percentage of customers corrected (solid blue line) in Fig. 9. This demonstrates the sensitivity of the comparison method to the configuration of customers which are mislabelled within the model. The proposed algorithm removes that sensitivity. Fig. 10 shows the overall accuracy for all customers for the same Monte Carlo simulation.

Using the proposed algorithm, the dependence on the utility phase labelling has been removed and this algorithm achieves excellent performance under the test conditions. At some point, with enough phase label errors in the utility model, the ability to assign predicted phases in the last step of the model will degrade. However, the clusters themselves will still represent correct phase clustering. It will simply be a matter of correctly assigning those clusters to the appropriate phase, and that could be done as post-processing by the utility.

### 5 Utility dataset results

#### 5.1 Dataset

The utility dataset used here comes from a utility operating in the North-Eastern United States and consists of ~1000 customers spanning ~15 months of data. The AMI data was collected using the 15-min averaged method. This dataset was also used in [14, 20]. Further analysis and characteristics of the dataset can be found there. Blakely et al. [20] give an overview perspective of common distribution system errors including phase identification. Fig. 11 shows a satellite image of the utility feeder. Based on [14], it is expected that the number of mislabelled phases is ~9%. Although there is no way to verify this without sending personnel into the

---

**Fig. 8** 25% of customers mislabelled confusion matrices. Red represents mislabelled customers and green represents correctly labelled customers.

**Fig. 9** Percent of mislabelled customers corrected.

**Fig. 10** Overall percent accuracy proposed method versus comparison method.
field, it is much less than would be required to make the algorithm fail in the final assignment step using the utility labels, as shown in Fig. 4. Therefore, the utility phase labels are used to make the final phase assignment. The following results reflect using that approach.

Validation of the results on the utility dataset is challenging as there are not ground truth labels for this feeder. Blakely et al. [14] present a two-stage validation methodology that combines validation using topology information and validation using publicly available Google Earth and Google Street View imagery. There are extensive details and images in [14] documenting the validation for that method. Although only a subset of customer phase labels can be verified due to issues such as availability of imagery, tree cover, underground cabling etc. there does exist a subset of customers for which the Google Earth imagery is reasonably conclusive for phase connectivity. Three novel examples using Google Earth satellite views and Street View are shown below as instances of the proposed algorithm making correct predictions.

### 5.2 Results

The proposed algorithm predicted that 143 customers were incorrectly labelled in the utility model of this feeder, out of a total of 1096 customers. This represents \( \sim 13\% \) of the customers. Those figures do include 41 customers which were labelled as being present on the feeder, but their phase was labelled as ‘unknown.’ Excluding those customers gives a prediction that 112 customers were incorrectly labelled out of 1055. This represents \( \sim 10.6\% \) of the customers. The inclusion or exclusion of those customers changes the results slightly for the other customers on the feeder. Following are two examples where the results of the proposed algorithm can be validated using Google Earth imagery.

Looking at Fig. 12, the left-hand image shows a transformer that was labelled in the utility model as Phase A but was predicted by the algorithm to be on Phase C. We can see the connection from the transformer connecting to the left-hand medium voltage wire at the top of the image. Continuing up the street, the next image shows the next transformer labelled as Phase C; this transformer is both labelled in the utility model and predicted by the algorithm as Phase C. However, it can be seen that this transformer connected to the same medium voltage line. This strongly indicates that this transformer represents an error in the original utility model and a corrected phase label predicted by the proposed algorithm.

Figs. 13 and 14 show another example of the proposed algorithm correcting errors within the original model. Fig. 13 shows the original labelling. All of the customers shown are labelled to be on Phase A which is served by the Phase A medium voltage line. Although not shown here for space reasons, continuing south on the street from the transformer in question, the next transformer which serves five customers is labelled (and predicted) as Phase C. This transformer is also connected to the same medium voltage line. This strongly indicates that this transformer represents an error in the original utility model and a corrected phase label predicted by the proposed algorithm.

### Table 1

| Max noise percentage | Number of incorrectly predicted customers | Percent accuracy, % |
|----------------------|------------------------------------------|---------------------|
| original             | 0                                        | 100                 |
| 0.05%                | 0                                        | 100                 |
| 0.15%                | 0                                        | 100                 |
| 0.25%                | 0                                        | 100                 |
| 0.35%                | 0                                        | 100                 |
| 0.45%                | 0                                        | 100                 |
| 0.55%                | 0                                        | 100                 |
| 0.65%                | 0.2                                      | 99.99               |
| 0.75%                | 0.4                                      | 99.97               |
| 0.85%                | 1.0                                      | 99.92               |
| 0.95%                | 3.6                                      | 99.69               |
| 1.0%                 | 7.6                                      | 99.46               |

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
Table 2  Confusion matrix results

| Feeder       | Utility label | A    | B    | C    | Total |
|--------------|---------------|------|------|------|-------|
| clustering predicted |              | 494,500 | 24, 24 | 7, 8 | 525,532 |
| A            | B             | 16, 16 | 224, 229 | 10, 3 | 250, 250 |
| B            | C             | 54, 48 | 10, 5  | 225, 222 | 289, 275 |
| C            | Total         | 564   | 258   | 233   |       |

The values in ‘italic font’ represent ‘proposed algorithm’; the values in ‘bold italic’ represent ‘comparison algorithm’.

Consider the ‘C’ row and column in Table 2. Three of the customers where the predictions differ are instances where the proposed algorithm agrees with the utility labels. The comparison algorithm predicted three customers which were originally labelled in the model as Phase C were Phase B. The proposed algorithm predicted that those three customers were actually on Phase C, as shown in the utility model. This can be seen in the addition of the three customers in row C, column C, from 222 customers in the comparison algorithm to 225 customers in the proposed algorithm. Analysis of those customers strongly suggests that incorrect labels in adjacent customers caused the comparison algorithm to misclassify those three customers and introduce error into the utility model while the proposed algorithm was able to correctly classify those customers. Analysis of the Google Earth imagery associated with these customers supports this conclusion.

This example is of a case where the phase predictions from the method presented here disagree with the predictions using the method in [14]. The Google Earth imagery clearly shows that this new method located and corrected a set of customers that remained as errors in the utility model. This method also maintained the correct label for a customer which was labelled correctly in the model which the comparison method predicted incorrectly. Fig. 15 shows an image with subfigures in each of the four quadrants. Home icons represent customers, light-blue hexagons represent transformers, red lines show Phase A lines, and blue lines show Phase C lines. The original labelling in the utility model is shown in the upper-left quadrant. In the upper-right image, showing the comparison algorithm predictions, two homes have changed phase. The one marked in yellow changed from C to A as did one marked in purple. The purple home that changed was a correct change. However, the customer marked in yellow was a correct label by the utility that was incorrectly predicted to be on a different phase. The yellow customer is actually on Phase C. The bottom quadrants are identical. In these quadrants, the proposed algorithm predictions and the actual labelling verified in Street View are shown. In addition to correctly labelling the customer marked in yellow, two customers to the right of this customer (orange) are also on Phase C, which is a change from A to C. Those two incorrectly labelled customers in the original model caused the comparison algorithm to misclassify the customer marked in yellow. Conversely, the proposed algorithm was able to determine the correct configuration.

It is worth noting that there are further examples of errors in this feeder that were corrected shown in [14]. Those error examples were also errors caught and corrected by the proposed algorithm. Other methods discussed in Section 2 have differing data requirements and assumptions and so are not directly compared here.

6 Discussion

Although the proposed algorithm compares favourably to the comparison algorithm, there are several aspects which have not been considered here. One additional aspect of complexity to be considered in future work is the differing phase connections. Three-phase customers were excluded from this analysis. Neither the synthetic dataset nor the utility dataset included delta configured phase connections. Both of those additions would add another level of complexity to the task. Additionally, the utility feeder in question did not include voltage regulators or capacitors.
Another aspect of future work includes testing on feeders under those conditions.

There is difficulty in directly comparing phase identification research and algorithms due to the variation in the available datasets as well as the variation in the characteristics of the datasets themselves such as the amount of data available, resolution, measurement interval, what data streams are available etc. One aspect of future work that is needed is a comparison of the proposed method with other methods which take an alternative approach to the phase identification task.

7 Conclusions

This paper proposed a method for distribution system model calibration based on clustering a coassociation matrix populated using an ensemble spectral clustering methodology. The proposed method achieves 100% accuracy on the synthetic dataset under rigorous testing conditions. It was shown to achieve superior performance when compared to a similar method from literature. It also removes the dependency on existing distribution system phase labels that have unknown quantities of error (or may not exist). Existing labels can be leveraged if they are considered trustworthy, or they may be excluded altogether. In that case, the algorithm returns a group of clusters that can be mapped to the correct phase as a post-processing step. On the utility dataset, ~10.6% of customers were predicted to be incorrect. Several examples of the algorithm correcting the phase label errors from the utility were shown, as well as examples of the proposed algorithm correctly classifying some customers incorrectly classified by the comparison algorithm. Overall, this method demonstrates excellent results on the phase identification task and merits further investigation under increasingly complex scenarios.

8 Acknowledgments

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 34226. 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. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

9 References

[1] Reno, M.J, Coogan, K., Grijalva, S., et al.: ‘PV interconnection risk analysis through distribution system impact signatures and feeder zones’, IEEE PES General Meeting, National Harbor, MD, USA, 2014

[2] Schwarz, V., Ghobabi, R.: ‘Transient over-voltage mitigation: explanation and mitigation options for inverter-based distributed generation projects’. Elect. Vehicle Transp. Cent., Sch. Ocean Earth Sci. technol., Univ. Hawaii Manoa, Honolulu, HI, USA, vol. Tech Report HNEI-02-15, February 2014

[3] Fairley, P.: ‘Hawaii’s solar push strains the grid’. MIT Technology Review, 2014, vol. NREL/TD-5500-57785, January 2015.

[4] Palmitzter, B., Broderick, R., Mathé, B., et al.: ‘On the path to sunshot: emerging issues and challenges in integrating solar with the distribution system’. National Renewable Energy Laboratory, 2016, vol. NREL/TP-5500-65311

[5] Nguyen, A., Velay, M., Schoene, J., et al.: ‘High PV penetration impacts on five local distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system aspects simulations’, Sol. Energy, 2016, 132, pp. 221–235

[6] Ebad, M., Grady, W.M.: ‘An approach for assessing high-penetration PV impact on distribution feeders’, Elect. Power Syst. Res., 2016, 133, pp. 347–354

[7] Luan, W., Peng, J., Maras, M., et al.: ‘Smart meter data analytics for distribution network connectivity verification’, IEEE Trans. Smart Grid, 2015, 6, p.1

[8] Cavraro, G., Aghanghadeh, R., Poulaki, K., et al.: ‘Data-driven approach for distribution network topology detection’. IEEE Power & Energy Society General Meeting, Denver, CO, USA, July 2015

[9] Peppanen, J., Reno, M.J., Thakkar, M., et al.: ‘Leveraging AMI data for distribution system model calibration and situational awareness’, IEEE Trans. Smart Grid, 2015, 6, (4), pp. 2050–2059

[10] Wong, V., Liao, W., Rajagopal, R.: ‘Distributed energy resources topology identification via graphical modeling’, IEEE Trans. Power Systems, 2017, 32, (4), pp. 2622–2694

[11] Gómez, A., Arguello, A., Quiros-Tortol, J., et al.: ‘Processing and correction of secondary system models in geographic information systems’, IEEE Trans. Ind. Inf., 2016, 15, (6), pp. 3482–3491

[12] Short, T.A: ‘A data-driven approach for phase identification, transformer identification, and secondary modeling’, IEEE Trans. Smart Grid, 2013, 4, (2), pp. 651–658

[13] Mitra, R., Kota, R., Bandypadhyay, S., et al.: ‘Voltage correlations in smart meter data’. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Sydney, Australia, 2015, pp. 1999–2008

[14] Blakely, L., Reno, M.J, Feng, W.: ‘Scalable clustering for customer phase identification using smart grid voltage timeliness’. IEEE Trans. Smart Grid, 2017, 1, pp. 1–24

[15] Lave, M., Reno, M.J., Broderick, R.J, et al.: ‘Full-scale demonstration of distribution system topology estimation to improve low-voltage circuit models’. IEEE Photovoltaic Specialists Conf. (PVSC), Washington, DC, USA, 2017

[16] Ashok, K., Reno, M.J., Osain, D., et al.: ‘Systematic study of data requirements and AMI capabilities for smart meter connectivity analytics’. IEEE 7th Int. Conf. on Smart Grid Engineering (SGE), Oshawa, ON, Canada, 2019, pp. 53–59

[17] Zhang, X., Grijalva, S.: ‘A data-driven approach for detection and estimation of residential PV installations’, IEEE Trans. Smart Grid, 2016, 7, (5), pp. 2477–2485

[18] Chen, D., Irwin, D.: ‘Sundance: black-box behind-the-meter solar disaggregation’. Proc. of the Eighth Int. Conf. on Future Energy Systems, Shatin, Hong Kong, May 2017, pp. 45–55

[19] Iyengar, S., Sharma, N., Irwin, D., et al.: ‘A cloud-based black box solar predictor for smart homes’. IEEE PES Trans. Cyber. Syst., 2017, 1, pp. 1–24

[20] Blakely, L., Reno, M.J, Peppanen, J.: ‘Identifying common errors in distribution system models’. Photovoltaic Specialists Conf. (PVSC), Chicago, IL, USA., June 2015, pp. 1–6

[21] Mohassel, R.R., Fung, A., Mohammadi, F., et al.: ‘A survey on advanced metering infrastructure’, Int. J. Electr. Power Energy Syst., 2014, 63, pp. 473–484

[22] ‘Advanced metering infrastructure and customer systems: results from the smart grid investment grant program’, US Department of Energy: Office of Electricity Delivery and Energy Reliability, September 2016

[23] Blakely, L., Reno, M.J, Ashok, K.: ‘AMI data quality and collection method consideration for improving the accuracy of distribution system models’. Photovoltaic Specialists Conf. (PVSC), Chicago, IL, USA, 2019

[24] Shepply, M., Brookhouse, N., et al.: ‘A novel method to bridge the gap supplied in the National Renewable Energy Laboratory's research support facility’. National Renewable Energy Laboratory (NREL), Technical Report – TP-5500-57785, April 2013

[25] Ma, Y., Fan, X., Tang, R., et al.: ‘Phase identification of smart meters by spectral clustering’. IEEE Conf. on Energy Internet and Energy System Integration (EI2), Beijing, China, October 2018

[26] Caïd, K.J.: ‘Meter phase identification’. US Patent 8,143,879, March 2012

[27] Achanta, S.V.: ‘Automatic phase identification for electric power delivery lines’. US 2019/038369 A1

[28] Wen, M., Arghandeh, R., Meier, J., et al.: ‘Phase identification in distribution networks with micro-synchrophasors’. Power and Energy Systems General Meeting (PES GM), Denver, CO, USA, 2015

[29] Sheinin, M., Schechner, Y., Kutulakos, K.N.: ‘Computational imaging on the electric grid’. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2363–2372

[30] Arya, V., Seetharaman, D., Kalyanaraman, S., et al.: ‘Phase identification in smart grids. IEEE Int. Conf. on Smart Grid Communications (SmartGridComm), Brussels, Belgium, 2011, pp. 25–30

[31] Pappu, S.J., Bhatt, N., Pamunthy, R., et al.: ‘Identifying topology of low voltage distribution networks based on smart meter data’. IEEE Trans. Smart Grid, 2018, 9, (5), pp. 5113–5122

[32] Xu, M., Li, R., Li, F.: ‘Phase identification with incomplete data’, IEEE Trans. Smart Grid, 2018, 9, (4), pp. 2777–2785

[33] Tang, X., Milanovic, J.V.: ‘Phase identification of LV distribution network with smart meter data’. IEEE Power & Energy Society General Meeting, Portland, OR, USA, 2018

[34] Foggo, B., Yu, N.: ‘On achieving supervised phase identification through the theory of information losses’, IEEE Trans. Smart Grid, 2020, 11, pp. 2337–2346

[35] Foggo, B., Yu, N.: ‘A comprehensive evaluation of supervised machine learning for the phase identification problem’, World Academy of Science, Eng. Technol. Int. J. Comput. Syst. Eng., 2018, 12, (6), pp. 419–427

[36] Pezeshki, H., Wolf, P.J.: ‘Consumer phase identification in a three phase unbalanced LV distribution network’. 2012 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, Germany, 2012, pp. 1–7

[37] Liao, Y., Wong, Y., Liu, G., et al.: ‘Unbalanced multi-phase distribution grid topology estimation and bus phase identification’, IET Smart Grid, 2019, 2, pp. 557–570

[38] Olivier, F., Sutera, A., Geurts, P., et al.: ‘Phase identification of smart meters by clustering voltage measurements’. Power Systems Computation Conf. (PSCC), Dublin, Ireland, 2018

[39] Wang, W., Yu, N., Foggo, B., et al.: ‘Phase identification in electric power distribution systems by clustering of smart meters data’. 2016 15th IEEE Int. Conf. on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016, pp. 259–265
[40] Wang, W., Yu, N.: ‘Maximum marginal likelihood estimation of phase connections in power distribution systems’. ArXiv (preprint), 2019

[41] Ng, A.Y., Jordan, M.I., Weiss, Y.: ‘On spectral clustering: analysis and an algorithm’. Proc. of the 14th Int. Conf. on Neural Information Processing Systems: Natural and Synthetic, Cambridge, MA, USA., December 2001, pp. 849–856

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: ‘Scikit-learn: machine learning in python’, J. Mach. Learn. Res. (JMLR), 2011, 12, (85), pp. 2825–2830

[43] Arthur, D., Vassilvitskii, S.: ‘k-means++: the advantages of careful seeing’. Symp. on Discrete Algorithms (SODA), New Orleans, LA, USA., 2007, pp. 1027–1035

[44] Yu, S.X., Shi, J.: ‘Multiclass spectral clustering’. Proc. of the Ninth IEEE Int. Conf. on Computer Vision (ICCV), Nice, France, 2003, vol. 2

[45] Vega-Pons, S., Ruiz-Shulcloper, J.: ‘A survey of clustering ensemble algorithms’, Int. J. Pattern Recogn., 2011, 25, (3), pp. 337–372

[46] Zhang, X., Jiao, L., Liu, F., et al.: ‘Spectral clustering ensemble applied to SAR image segmentation’, IEEE Trans. Geosci. Remote Sens., 2008, 46, (7), pp. 2126–2136

[47] Fuller, J., Kersting, W., Dugan, R., et al.: ‘Distribution test feeders’. IEEE PES AMPS DSAS Test Feeder Working Group, 2013. Available at http://sites.ieee.org/pes-testfeeders/

[48] ‘American National Standards Institute’, 2015. Available at http://www.nema.org/stds/c12-20.cfm