SMART-METER PHASE CLUSTERING IN LOW-VOLTAGE DISTRIBUTION NETWORK

Pedro Carriço*, Adriana Leal, Daniela Jordão, Bruno Galhardo

ENEIDA.IO, Coimbra, Portugal
**pcarrico@eneida.io*

Keywords: PHASE IDENTIFICATION, SMART METERS, VOLTAGE TIME SERIES, CLUSTERING, GRID MAPPING

Abstract

Low-voltage (LV) distribution networks have evolved over decades with limited prioritization of detailed documentation and digitization. Unlike high-voltage (HV) networks, which are well-mapped and monitored, LV networks often lack critical data, such as labelled feeders and phase identification in GIS systems. This data gap combined with the growing integration of distributed energy resources (DER) and the adoption of time-of-use tariffs presents challenges for effective planning and operations.

This paper introduces a method for phase identification based on voltage measurements from smart meters (SMs). The analysis utilizes a real-world dataset comprising 15-minute sampled voltage time series from 89 and 109 single-phase SMs connected to two transformers. The developed machine learning pipeline thoroughly addresses data pre-processing and applies spectral clustering and unsupervised clustering evaluation. The pipeline was designed to ensure robustness against distinct voltage profiles often observed for different transformers.

By utilizing existing SM data, the proposed approach eliminates the need for costly on-site data collection. It enables distribution system operators to automatically map phase connectivity, delivering immediate benefits such as improved load balancing, loss reduction, and enhanced network planning. This method provides a scalable, cost-effective solution for addressing critical data gaps in LV distribution networks.

1 Introduction

The low-voltage (LV) distribution network faces increasing operational challenges due to the growing integration of distributed energy resources (DERs) and the adoption of new smart loads like electric vehicles and heat pumps [1, 2]. Efficient network management and planning requires distribution system operators (DSOs) to have accurate knowledge of their network topology, particularly regarding customer phase connectivity. Without accurate phase grid mapping, DSOs cannot effectively assess three-phase imbalances, compute line losses, or optimize network operations [3, 4].

However, in many LV distribution networks, the phase connectivity information is often missing, outdated or incorrect. This information gap typically arises from poor historical record-keeping, network reconfigurations during maintenance, and the lack of systematic documentation during initial customer connections [1, 3, 4]. While traditional methods for phase identification exist, such as manual field measurements or signal injection techniques, these approaches are time-intensive, costly, and impractical for large-scale deployment [2, 5].

The widespread deployment of smart meters (SMs) presents an opportunity to develop data-driven approaches for phase identification by collectively providing a considerable amount of voltage measurements, enabling methods that leverage patterns in this data to infer phase connections [5]. Most studies present voltage-based phase identification solutions that consist in assigning customer phase according to the highest similarity with each substation feeder phase (customer-to-feeder similarity). However, in the absence of substation feeder phase time series, there are methods exploring customer-to-customer voltage similarity [3, 4]. In this scenario, it is assumed that there are impedances and electric loads on the distribution system that, in turn, lead to unbalanced line currents and voltages. As result, observing high similarity among voltage time series in the same phase is expected [3, 4, 5]. Additionally, customers connected to different distribution transformers are expected to exhibit different voltage profiles [4].

Several approaches have been proposed for phase identification using voltage measurements. Interestingly, when comparing voltage with the alternative power measurement-based phase identification methods, results show that voltage methods lead to better performance [2]. Summarily, voltage-based costumer-to-customer methods that do not rely on available transformer/feeder time series [5, 6, 7, 8, 9], use correlation [9, 8], *k*-nearest neighbour [7] similarity matrices to which spectral clustering [5, 6, 7, 9] or constrained k-means [6, 8] is applied. Reported constrained clustering relies on information from three-phase meters [8] or the customer-transformer labelling [6]. Some of these studies [5, 7] are also conducted on synthetic datasets. Additionally, results are often reported considering the existence of ground truth for the phases corresponding to each costumer/SM. In

other words, the proposed methods are assessed using supervised (ground truth-based) metrics and lack the unsupervised assessment of the clustering solutions which are required when ground truth information is not available and different clustering methods are tested on the training data.
Based on the above, several technical challenges need to be addressed in phase identification problem:

1. Data quality: SM measurements may contain noise, missing values, or inconsistencies that need to be handled appropriately.
2. Voltage profile variations: different transformers can exhibit distinct voltage profiles, requiring methods that are robust across diverse network conditions.
3. Performance validation: the lack of ground truth data makes it challenging to evaluate clustering accuracy.
4. Scenario flexibility: deriving a method robust to different acquisition scenarios.

In this paper, we present a data-driven unsupervised method for phase grid mapping using SM voltage data. The method employs spectral clustering techniques which are particularly well-suited for capturing the inherent structure in voltage measurements while being robust to noise. Since voltage profiles can vary significantly between different transformers due to diverse network characteristics and loading conditions [10], our method uses transformer-customer labelling to provide transformer-specific phase identification. This way, spectral clustering model can capture the unique voltage patterns and relationships within each transformer's service area. As shown by Hoogsteyn et al. [2], voltage measurements from customers connected to the same transformer tend to exhibit similar temporal patterns that differ from those connected to other transformers.
Furthermore, network reconfigurations and maintenance activities can occur periodically, suggesting the need for regular batch analysis to detect such changes and inform the DSO.

The key contributions of this work include:

1. A comprehensive pre-processing pipeline to handle missing values and prepare SM data for clustering analysis.
2. A systematic evaluation of clustering parameters and model performance using unsupervised clustering evaluation metrics.
3. Validation of the method's robustness across multiple distribution networks with distinct voltage characteristics.

The remainder of this paper is organized as follows. Section 2 describes the dataset and methodology, including data pre-processing and the clustering approach. Section 3 presents and discusses the results. Finally, conclusions and future work are presented in Section 4.

2 Methodology

The following sections describe each step of the machine learning pipeline designed for this study (see Fig. 1).

2.1 Smart meter data

In this study, we used a dataset comprising 15-minute sampled voltage time series for 89 and 109 single-phase SMs connected to two transformers A, and B, respectively. These transformers are part of a real LV grid in Évora, Portugal, operated by DSO E-Redes [11]. The dataset contains phase labels for each SM obtained using a phase identification approach based on substation data [12]. The voltage time series include a period of 45 days spanning across the months of April and May. These data are then first grouped by transformer ID and then split by date to yield two-weeks long datasets. For each transformer, we are left with three two-weeks datasets: “Batch 0” with data from April, and “Batch 1” and “Batch 2” with data from May. Each of these will be independently pre-processed and clustered afterwards. Analysis of two-week batches provides an optimal balance between data volume and processing efficiency while allowing detection of systematic changes in phase connectivity.
Fig. 2 presents the distribution of the mapped SM phases per transformer. The distribution of phases for Transformer A indicates that the distribution is unbalanced, since 47% of the SM are in phase C. This unbalance can lead to bias during the phase identification method for this transformer, since one phase is more predominant than the others in the dataset [13, 14]. For Transformer B, the SM are evenly distributed across all phases.

2.2 Data pre-processing

Since the data obtained by the SMs may contain outliers and/or missing values, we added a pre-processing step to ensure that quality and robustness of the data available was the best possible. The pre-processing starts with a filtering stage. At this stage, SMs with more than 70% of missing data and SMs with constant data are removed. In addition, the zero values in the voltage series are considered as missing values, since we could not guarantee that they were not outliers. Fig. 3 shows the distribution of missing data in the three batches.

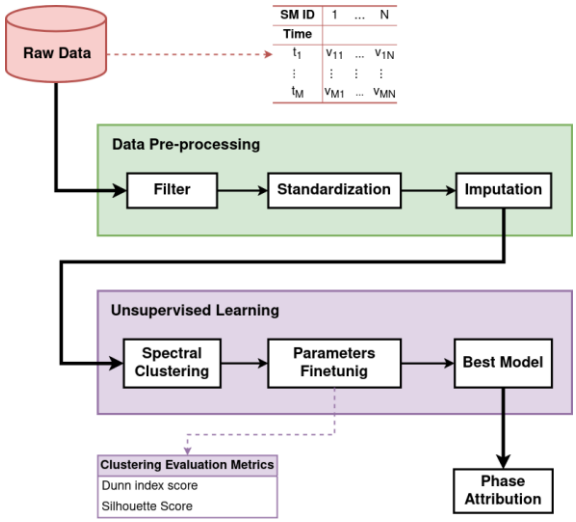

Fig. 1 Methodology flowchart depicting the machine learning pipeline developed for phase identification using voltage measurements. This methodology was applied independently for each transformer/batch data.

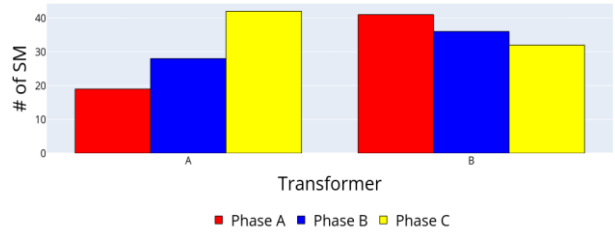

Fig. 2 Distribution of the real phases of the SM for all datasets.

The voltage time series contain missing values resulting from either hardware issues or problems related to the infrastructure. The distribution of missing data indicates that most SM have a percentage of missing data below 10%.
In the next stage of pre-processing, the dataset is scaled to ensure that the comparisons between the voltage time series are not affected by large variations in amplitude and offset [15, 16]. We choose the Scikit-Learn Standard Scaler since we have previously filtered our datasets to ensure that the outlier voltage values that may have been present in data from a real LV grid were dealt with [16].

The last step of the pre-processing stage is the imputation of missing values. An essential property of a robust time series dataset is a low or non-existent percentage of missing values, i.e., a dataset with continuous data, since it helps to obtain more accurate models and predictions [17]. It is important to note that time series, contrary to other types of data, have temporal dependence [15]. Therefore, we used a multivariate imputation algorithm, the Scikit-Learn KNN Imputer method, to perform the imputation of missing values.
Since no SM were removed during the pre-processing stage, the distribution of real phase of the SM did not change from the one presented in Fig. 2.

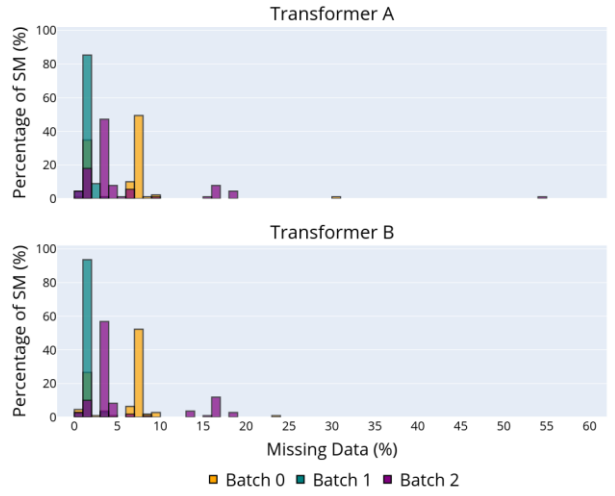

Fig. 3 Distribution of missing values for Transformer A and Transformer B for the three batches.

2.3 Spectral clustering
Spectral clustering is an unsupervised machine learning technique that leverages graph theory principles to perform dimensionality reduction before clustering. This approach is particularly well-suited for phase identification as it can capture the inherent non-linear relationships in voltage measurements while being robust to noise.
The algorithm consists of three main steps:

1. Construct a pairwise similarity matrix, representing the relationships between SM voltage time series. We tested different methods to compute the similarity between each pair of SMs: *k*-nearest neighbour graph (for different values of nearest neighbours), radial basis function Gaussian kernel with Euclidean distance (for different values of gamma) and Pearson's correlation coefficient (option "precomputed affinity matrix").
2. Compute the normalized graph Laplacian from the similarity matrix.
3. Map the data into a nonlinear-dimensional space using the eigenvectors of the Laplacian graph corresponding to the *k*-smallest eigenvalues (where *k=3* for our three-phase identification problem).
4. Finally, different clustering methods are applied in this new embedding space to assign phase labels: k-means, discretization ("discretize" option) and simple column-pivoted QR factorization ("cluster_qr" option).

The options described previously are the choices provided by Scikit-Learn implementation of spectral clustering.
Spectral clustering offers several advantages for phase identification:

- It can identify clusters of arbitrary shape, not just spherical ones.
- It reduces the dimensionality of the voltage time series data while preserving essential phase relationships.
- The spectral embedding helps separate the data in a way that makes the subsequent clustering more effective.
- It is less sensitive to noise and outliers compared to direct clustering of the voltage measurements.

2.4 Clustering evaluation
The clustering results are evaluated using unsupervised metrics to assess the quality of the phase assignments without requiring ground truth labels. The following unsupervised metrics were considered: silhouette score and Dunn's index. These metrics collectively provide a comprehensive evaluation of the clustering quality, examining both the compactness of phase assignments and the separation between different phases. Their complementary nature helps validate the robustness of the phase identification results from different analytical perspectives. Additionally, as we have knowledge about real phases of each SM (ground truth), we also computed a supervised clustering evaluation metric: adjusted rand index (ARI), only for the purpose of validating the results for this specific dataset for which this information is available. ARI can be considered a measure of accuracy in the sense that it measures the similarity between the real labels and the clustering labels, by comparing pairs of elements and quantifying whether these are grouped together or separately. Importantly, ARI is invariant to cluster label permutations,

which is important in this application scenario as the clustering might assign a given cluster as "1" and the same number of points is assigned "0" in the real labels [18].

2.5 Best model selection based on unsupervised metrics

The best clustering method/parameters are selected through a two-stage filtering process that considers both cluster stability and quality. The silhouette score and the Dunn's index are used to identify stable and well-defined clustering solutions. Models are filtered based on two criteria:

1. Stability criterion: the absolute difference in silhouette and Dunn's index scores (independently) between consecutive parameter configurations should be less than 0.05, indicating a stable solution that is not overly sensitive to parameter changes.
2. Quality threshold: stable values of the unsupervised scores were obtained for values of nearest neighbours higher than 13 in spectral clustering based on k-nearest neighbour graph.

If no solutions meet both criteria, all models are considered in the next stage. Among the filtered models (or all models if no solutions meet the initial criteria), the final selection corresponds to the maximum of Dunn's index and silhouette scores.

3 Results

The phase identification analysis was performed independently for each transformer and batch period. Fig. 4 shows the similarity matrix for transformer A's April dataset (batch 0), where clear block diagonal patterns indicate strong within-phase correlations, suggesting natural clustering of voltage profiles by phase.

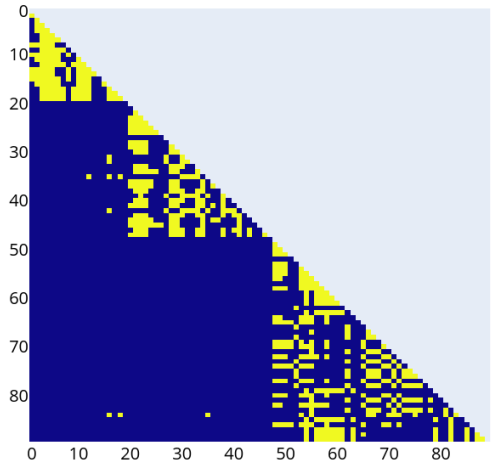

Fig. 4 Heatmap depicting the binary similarity matrix obtained for transformer A, for the April data (batch 0). When a pair of SMs voltage time series is considered similar by the *k*-nearest neighbours graph method (for 14 neighbours), the pair is assigned 1 in the binary metric (yellow) and 0 otherwise (blue). The matrix was sorted by each SM real phase for visualisation purposes.

Fig. 5 presents the clustering performance of the best models selected, using the method described in section 2.5, for the different batches and transformers. By looking at the results, it is possible to conclude that both transformers show high phase identification accuracy for transformers A and B and most of the batches, using spectral clustering based on *k*-nearest neighbour graph. Different clustering methods (k-means, discretize, cluster_qr) performed optimally for different batches, suggesting that a flexible approach to method selection is beneficial. Batch 2 shows particularly low performance for transformer B (ARI = 0.45), indicating that temporal variations in voltage time series were not specific enough to allow distinguish phases. In fact, visual inspection of the SMs time series reveals two SMs profiles that diverge from the mean SM profile.

The results demonstrate that batch transformer-specific analysis is beneficial, as the unsupervised metrics performance differs among the six datasets, supporting our approach of independent transformer/batch analysis.

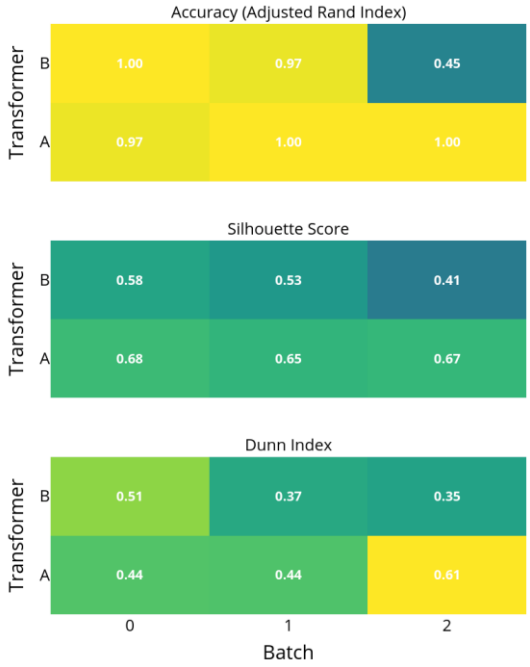

Fig. 5 Performance metrics of the best models across the different transformers and batches.

4 Conclusion

This study presents a robust methodology for phase grid mapping in LV networks using SM voltage measurements. The proposed spectral clustering approach achieves high performance in phase identification, with accuracy exceeding 0.9 in optimal conditions. The method underperforms for a given period and transformer, highlighting the importance of batch processing and transformer-specific analysis in identifying problems in the grid. The unsupervised evaluation framework successfully identifies optimal clustering parameters without requiring ground truth phase information, making it applicable to realistic scenarios where phase

information is often unknown. Future work entails validating that cluster (phase) membership of SMs does not change over defined periods, investigating the causes of performance variations in different batches, and extending the methodology to handle dynamic network reconfigurations. Additionally, exploring the impact of different voltage measurement sampling rates and longer time periods could provide insights into optimal data collection strategies for phase identification. The method's success in identifying phases using only voltage measurements makes it a valuable tool for DSOs seeking to improve their network topology information, particularly in areas where traditional phase identification methods are impractical or costly.

5 Acknowledgements

We acknowledge E-Redes for providing the dataset used to develop this study. The work presented in this paper is currently undergoing further development at ENEIDA.IO and is supported by the ATE - Alliance for the Energy Transition (project code C644914747-00000023, agenda 56) co-funded by the Recovery and Resilience Plan (PRR) through the European Union.

6 References

[1] Al-Jaafreh M. A. A., Mokryani G.: 'Planning and operation of LV distribution networks: a comprehensive review' IET Energy Syst. Integr., 2019, 1, (3), pp. 133–146

[2] Hoogsteyn A., Vanin M., Koirala A., et al.: 'Low voltage customer phase identification methods based on smart meter data' Electr. Power Syst. Res., 2022, 212, pp. 108524

[3] Wang W., Yu N., Foggo B., et al.: 'Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data' 2016 15th IEEE International Conference on Machine Learning and Applications, pp. 259-265

[4] García S., Mora-Merchán J. M., Larios D. F., et al.: 'Phase topology identification in low-voltage distribution networks: A Bayesian approach', Int. J. Electr. Power Energy Syst., 2023, 144, pp. 108525

[5] Blakely L., Reno M. J.: 'Phase identification using co-association matrix ensemble clustering', IET Smart Grid, 2020, 3, (4), pp. 490-499

[6] Wang W., Yu N., Foggo B., et al.: 'Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data', 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 259-265

[7] Ma Y., Fan X., Tang R., et al.: 'Phase Identification of Smart Meters by Spectral Clustering', 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), pp. 1-5

[8] Olivier F., Sutera A., Geurts P., et al.: 'Phase Identification of Smart Meters by Clustering Voltage Measurements', 2018 Power Systems Computation Conference (PSCC), pp. 1-8

[9] Blakely L., Reno M. J., Feng W.: 'Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries', 2019 IEEE Power and Energy Conference at Illinois (PECI), pp. 1-7

[10] Zhou L., Zhang Y., Liu S., et al.: 'Consumer phase identification in low-voltage distribution network considering vacant users', Int. J. Electr. Power Energy Syst., 2020, 121, pp. 106079.

[11] Puertas de la Morena, A., Santos, R., Lópes, D., et al.: 'Integrated monitoring system to assess LV flexibility impact', CIRED Open Access Proc. J., 2020, vol. 2020, (1), pp. 140-142.

[12] Galhardo, B., Cordeiro, F., Cerqueira, T.: 'Energy box mapping and energy balance in low voltage grids', Proc. 26th Int. Conf. and Exhibition on Electricity Distribution (CIRED), Online Conference, 2021, pp. 2646-2649.

[13] Qian, J., Saligrama, J.: 'Spectral clustering with imbalanced data', 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 2789-2793.

[14] Aksoylar, C., Qian, J., Saligrama V.: 'Clustering and Commuting Detection with Imbalanced Clusters', IEEE Trans. Signal Inf. Process. Netw., 2017, vol. 3, (1), pp. 61-76.

[15] Lima, F. T., Souza, V. M. A.: 'A Large Comparison of Normalization Methods on Time Series', Big Data Res., 2023, 34, (100407), pp. 1-8.

[16] Amorim, L. B. V. de, Cavalcanti, G. D. C., Cruz, R. M. O.: 'The choice of scaling technique matters for classification performance', Appl. Soft Comput., 2023, 133, pp. 109924.

[17] Zainuddin, A., Hairuddin, M. A., Yassin, Z. I.A. et al.: 'Time series data and recent imputation techniques for missing data: A review', Proceedings of the 2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri Sarawak, Malaysia, pp. 346-350.

[18] Warrens, M.J. and van der Hoef, H. 'Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs' *J. Classif., 2022,* 39, (3), pp. 487–509.