21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE US National Committee
2022 Grid of the Future Symposium

Data-Driven Scalable Phase Identification in Modern Distribution Grids

B. SOHRABI, A. KHODAEI
University of Denver
USA

SUMMARY

Power grid operation and maintenance procedures rely on precise grid topology. However, the connectivity of the components is often altered with unpredictable and unavoidable modifications throughout the grid's lifetime. As a result, accurate information about consumer connectivity is not always available. Identifying the phase connectivity of consumers is one of the first steps in reconstructing the grid topology data. There are multiple data-driven solutions to answer the phase identification problem, which are commonly faster and less expensive than the hardware-based and fieldwork-oriented alternatives. One of the disadvantages of these methods is the inability to solve the problem accurately when the number of loads in the network increases and when real-world scenario problems such as measurement and distribution losses are applied. This paper aims to demonstrate a data-driven approach to answer the scalability issue of phase identification methods. In this paper, the smart meter and substation measurements of phases in a low-voltage distribution network are utilized to identify consumers' connectivity to corresponding phases and reassemble the topology data. Furthermore, the effect of measurement loss and distribution loss is simulated, and a set of methods are presented to address the scalability issue of the proposed optimization model. The model is then tested with a dataset with over five thousand real load profiles with different amounts of simulated losses. In addition, the effect of lower Advanced Metering Infrastructure (AMI) penetration is considered in the model. The proposed method is benchmarked in different scenarios, showing a satisfactory performance to solve the problem for large-scale grids.

KEYWORDS

Advanced metering infrastructure, Smart meters, Network topology, Phase identification, Machine learning, Data analysis

Behrouz.Sohrabi@du.edu

1. INTRODUCTION

Outage detection, load balancing, and demand forecasting, some of the most routine power grid operation and maintenance procedures amongst many others, rely on precise grid topology data [1]. However, the grid topology is often altered with unpredictable and unavoidable modifications, e.g., through grid restoration, that often are not well documented [2]. The frequency and dispersion of these changes during a period when certain portions of the network are under emergency conditions are one of the reasons they are not documented. As a result, the accuracy of the initial topology data will decrease, and reliable data on loads and consumer connectivity may not always be accessible. Identifying the phase connectivity of consumers is one of the first steps in reconstructing the topology data.

Primarily there are two techniques for phase identification. One of these techniques involves adding additional hardware to the network, which requires field operations [3], [4]. This method is costly and challenging to put into practice. Using merely data is the other technique utilized for the phase identification problem. These data must be either retrieved or readily accessible by installed AMI to the utilities. An increasing number of utilities retrieve and store AMI voltage measurements, however, there may still be electric utilities that do not store the voltage data. [5]. Therefore, using the voltage measurement on a broad scale could be inapplicable owing to infrastructure constraints, and if they are accessible, they could have restricted coverage both in terms of volume and time.

Consumption measurements in watts are the data continuously provided by smart meters [6], gathered by data collectors, and transmitted to the utility's data centers. These data are the most available data from low-voltage distribution networks. The resolution of these measurements is typically 60, 30, 15, or 5 minutes [7].

The implementation of data-driven phase identification techniques in low-voltage distribution networks is based on statistical, machine learning, and optimization techniques. Relatively more accurate techniques that exploit voltage data try to infer which load is closer to which phase in terms of correlation [8], [9]. It is done by calculating the linear correlation coefficient of the measured low-voltage distribution network load voltages with the voltages of each phase in the medium-voltage distribution network at specific times of a time-series dataset. Alternately, they attempt to cluster the load voltage time series utilizing a machine-learning algorithm, assigning each cluster to a certain phase [10]. Other than the unavailability of this specific type of data, several issues with these techniques include the inaccuracy of time synchronization, voltage profile similarity, and medium-low voltage conversion accuracy. Correlation [11], graph theory [12], and modeling through mixed-integer programming (MIP) [13] are other methods that use data from consumption measurements. Also suggested are ensemble approaches [14] that use both consumption and voltage measurements. As consumers' consumption patterns are, by nature, similar, correlation-based methods cannot be applied to real-world scenarios. However, despite their remarkably high accuracy, MIP methods that use the smart meters’ load profiles, encounter issues such as exponential growth of problem complexity as the number of loads and accordingly measurements increase. The proposed data-driven techniques that have been presented up to this point often involve synthetic data [13] or minimal populations and sample sizes [15], [16] for evaluation and benchmarking.

Nevertheless, MIP techniques that employ consumption measurements are one of the most precise and efficient data-driven techniques for tackling phase identification problems. In this article, we will propose data analysis and machine learning-based approaches for scaling this technique and bringing it closer to solving real-world problems.

2. PROPOSED SOLUTION

Assuming that the amount of power transferred through the distribution substation and the amount of consumption by each phase at specific time intervals in a grid can be measured, the difference, which represents the distribution network's total loss, is calculated as follows:

$$L = \sum_{p=1}^{3} \sum_{t=1}^{\mathrm{T}} L_{pt} = \sum_{p=1}^{3} \sum_{t=1}^{\mathrm{T}} (D_{pt} - C_{pt}) \tag{1}$$

Where D and C represent the value of supplied and consumed power at each phase, respectively. The consumed power can be decomposed into two matrices. The first matrix (A) contains records of smart meter consumption measurements, while the second matrix (X) denotes the connection of each smart meter to the corresponding phase.

$$C_{pt} = A_t \, . X_p \quad \forall p, \forall t \tag{2}$$

We propose the following optimization model to calculate the unknown binary variables while accounting for distribution loss in (1) for the measurement data at various time intervals.

$$\min_{x} \sum_{p=1}^{3} \sum_{t=1}^{T} (D_{pt} - L_{pt} - \sum_{n=1}^{N} (a_{tn} x_{pn})) \tag{3}$$

$$\text{s.t.} \quad x_{pn} \in \{0,1\} \qquad \forall p,\ \forall n \tag{4}$$

$$\sum_{p=1}^{3} x_{pn} = 1 \quad \forall n \tag{5}$$

where x is a binary value for each smart meter n ($n \in N$), indicating its connectivity to phase p, and a is an individual measurement in the matrix A representing each measurement of smart meter n during the time interval t. Since the loss value is always positive and the constraints and variables of the objective function are linear, this optimization problem can be addressed via mixed-integer linear programming (MILP). This model can effectively identify the phase connectivity of each meter. However, it is not scalable and can only be used with a limited number of meters. We propose the following five approaches to enhance the accuracy and scalability of the proposed model, as discussed in the following.

A. Smart Meter Clustering

During the lifetime of a distribution network, not all lines and loads may be altered, and the initial topology data available to the utility will not become completely invalid. Therefore, the assumption of total network unobservability is not a reasonable approach. From the initial topology data, the information may be extracted to considerably aid in simplifying and reducing the number of model parameters, while not violating the method's underlying assumption that the phase to which the smart meter is assigned is unknown. We propose a machine learning-based approach using Density-based spatial clustering of applications with noise (DBSCAN) algorithm for extracting two types of information from a small proportion of topology data.

1. Identifying the smart meters on the same phase (regardless of the exact phase they are on). It will result in some of the smart meters' measurements being grouped together, lowering the number of problem variables.

2. Identifying the smart meters in different phases. This information can further be fed into the MILP model by specifying new constraints limiting the feasible region and ultimately reducing calculation time. The new constraints will be added to (1) as follows:

$$\sum_{m=1}^{M} x_{pm} < B \ \ \forall p, M \ \subset N \tag{6}$$

where B represents the total number of smart meters identified as being on different phases.

B. Profile’s Similarity Identification

Since the solver is essentially attempting to derive the value of supplied power matrix (D) from the values of the consumption measurements matrix (A) by adjusting the binary values in the connectivity matrix (X), if the values in each column of A or D are equal or very close, the values of X may be incorrectly determined, and the solver must take an additional step to optimize the problem. Considering this, data points might be selected more optimally based on a similarity score (S) derived from the standard deviation of measurements (σ_t) as follows:

$$\sigma_t = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_{ti} - \mu_t)^2} \tag{7}$$

$$S_t = 1 - \frac{\sigma_t - min\,(\sigma)}{max(\sigma) - min\,(\sigma)} \tag{8}$$

Having the similarity score for all data points in the time series, those with values above a certain threshold can be discarded. Typically, the period of measurements that utilities have access to exceeds the minimum window necessary to address the problem and applying a filter on the given dataset may not result in a substantial loss. However, using the most recent measurement values and the appropriate data point filtering criteria is the more rational course of action.

C. Measurement Loss Considerations

Another factor may play a role in selecting or filtering data points that contribute to the MILP model’s simplification and reduction in the number of parameters. Not all smart meters are in close range of the data collectors; those further from the collectors must transmit measurements to smart meters in closer proximity. Occasionally, the communication network may not function as intended and data may be lost. Thus, the data is highly likely to contain loss values. The loss does not necessarily have a uniform distribution. Therefore, considering the loss values' effect necessitates an analysis of the datasets already recorded. This loss causes the values of supplied power matrix and the values of the consumption measurements matrix to diverge in the time-series data points (columns in D and A matrices). If many measurements in each column of A are zero, solving the MILP model in that specific time will not result in the correct solution and will merely add to the problem's parameters and complexity. In addition, discarding entire data points when certain smart meter values are zero will result in the loss of crucial information. The ideal case is when zero values are limited to smart meters in one or two phases. In that case, the problem for the remaining phase(s) without a zero value would be resolved. Due to the uncertain phase assignment of smart meters, it is impossible to be optimistic about this case. Hence, different situations of measurement losses should be evaluated with different thresholds to determine the most effective threshold for discarding data points with zero values.

D. Distribution Loss Considerations

The necessity to accurately model the distribution loss is another scalability limitation of this technique. From 2016 to 2020, the distribution loss is estimated to be around 5% in the United States [17]. This value depends on various technical and non-technical factors and may not have a particular and stable value. Since the amount of energy loss has a quadratic relationship with the intensity of the current, it can be said that this amount will be higher during peak consumption hours. Consequently, attempting to represent this variable with a specific percentage won't be realistic and will lead to inaccuracy when solving the problem under actual circumstances and with actual data. In this research, an approach based on machine learning and time series prediction is utilized to model the distribution loss variable optimally so that it can be incorporated into the optimization model.

E. Solution Optimality Gap Relaxation

With more load profiles, the minimum number of measurements required to solve the above problem grows, so the solver needs more variables to arrive at the optimal solution. In this case, the probability that each phase's load measurement and each smart meter's load measurement will be approximately the same at each data point escalates. As a result, the complexity of the objective function (3) increases exponentially with the number of smart meters (N) and, consequently, the total number of time intervals (T). By relaxing the gap tolerance criterion, the solver will terminate the optimization process when it reaches an optimal integer solution for the objective value with respect to the optimal value within the feasible region. As a result, by selecting the optimal gap tolerance value, the solver will need less time to reach an accurate solution, which contributes to the scalability of this technique.

3. NUMERICAL RESULTS

The dataset [18] used in this section contains over five thousand smart meter consumption measurements with an interval of 30 minutes over four years. In each step, different scenarios are tested on this dataset through sensitivity analysis to find the best parameters for the solver.

Case 1. Finding the optimal number of measurements for 150 smart meters: The model is used to classify 150 smart meters with different values of gap tolerance. The maximum computation time is set to 600 seconds. The results demonstrate a tradeoff between the computation time and the model's accuracy. In general, the optimal value of the gap tolerance increases with the number of measurements and there is a point where the computation time drops with increasing the measurement window. This could be considered the optimal gap tolerance for that sample size. For the studied case, the optimal gap tolerance is 5% with 192 measurements, as shown in Figure 1.

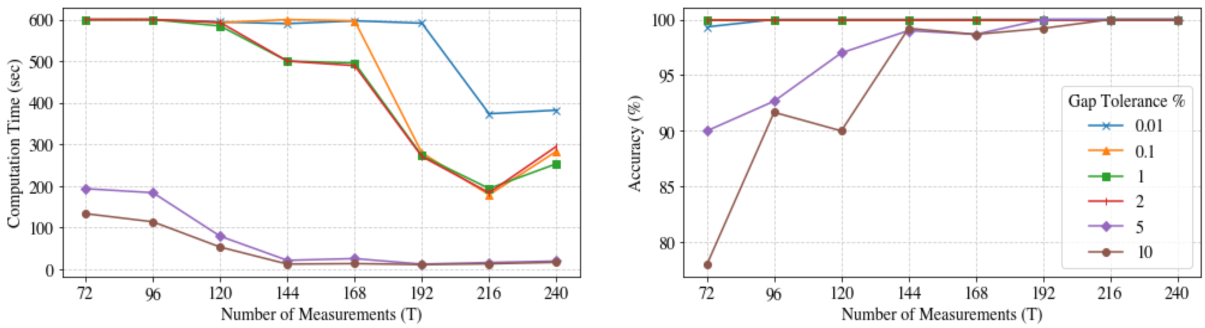

Figure 1. Impact of gap tolerance relaxation on computation time, accuracy, and the number of measurements (T)

Case 2. The Impact of Distribution Loss: Using a fixed gap tolerance, the model is tested over a sample size of 200 meters with three scenarios for modeling the distribution loss, as shown in Figure 2. Without considering the distribution loss, the model would need more time and measurements to achieve high accuracy. Using a fixed percentage constant to model the loss from the amount of total energy shows a better performance. Modeling the distribution loss with a machine learning regressor algorithm and then feeding it to the model's formulation as a variable value for each data point is the best approach.

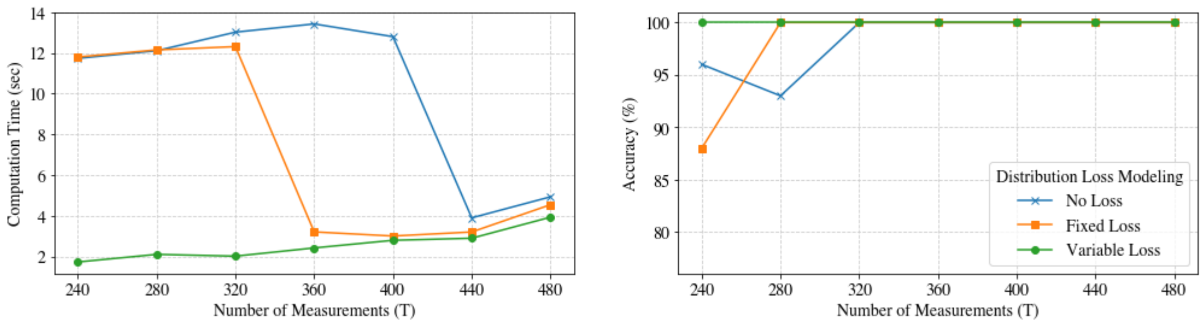

Figure 2. Performance of modeling distribution loss for 200 smart meters

Case 3. The Impact of Measurement Loss: Through a sensitivity analysis, different percentages of loss are simulated and tested on 200 and 300 smart meters when the data points with different percentages of zero values are discarded (Table 1). The model's accuracy is 100% in every scenario, and the best case is determined with the least computation time. This analysis shows that when the number of available measurements is limited, choosing not to discard all the measurements containing zero values will increase the performance in terms of computation time.

Table 1. Performance in the presence of measurement loss with various discarding thresholds

N	T	Zero Values	Completeness Threshold	Computation Time (s)	Accuracy
200	240	10%	0%	4.339	100%
			20%	4.608	
			50%	**4.016**	
			100%	4.086	
	360	30%	0%	5.660	
			20%	5.535	
			50%	6.308	
			100%	**4.444**	

	480	50%	0%	7.406	
			20%	8.103	
			50%	7.028	
			100%	**4.615**	
300	360	10%	0%	15.230	100%
			20%	14.217	
			50%	**11.816**	
			100%	11.885	
	480	30%	0%	17.195	
			20%	17.747	
			50%	13.655	
			100%	**10.792**	
	600	50%	0%	18.650	
			20%	16.953	
			50%	14.462	
			100%	**10.327**	

Case 4. Considering the profiles' similarity: Figure 3 display the data points for which the similarity score of the supplied power matrix for each phase (left) and the consumption measurements (right) exceed a 0.7 threshold, which typically occurs during off-peak hours. Having the similarity score for these two values in each measurement, the model is tested with different levels of discarding threshold with Monte Carlo simulation, ran 50 times on 500 and 1000 sample sizes. The results (Table 2) indicate that discarding more measurements will increase the accuracy in exchange for increasing the computation time.

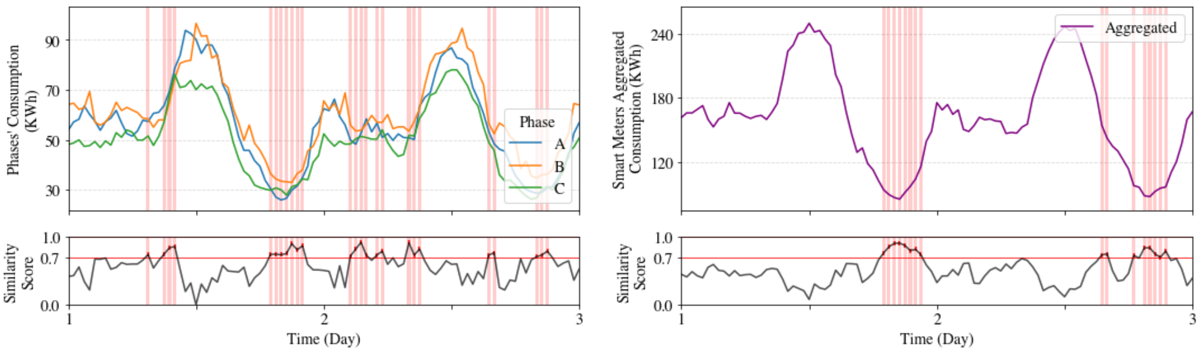

Figure 3. Similarity score over 96 measurements (2 days) of 500 smart meters

Table 2. Performance of the model after and before removing data points with a high similarity score

N	T	Similarity Discard Threshold	T After discarding	Average Computation Time (s)	Average Accuracy
500	1440	~ 0.7	480	28.248	100%
	720	~ 0.8		28.109	100%
	480	Without Discarding		22.796	99.6%
1000	2400	~ 0.7	960	235.190	100%
	1680	~ 0.8		249.672	98.2%

Case 5. Clustering smart meters: This approach will reduce the dimensionality of the model by considering the measurements of smart meters in a cluster as one measurement, and accordingly decreases the overall computation time as well as the required number of measurements (T). In this case, single-phase consumers are grouped using the topology data of a 2500-Bus test network. The model is tested in two scenarios with 218 (Figure 4) and 109

clusters generated by the DBSCAN model. Computation time dropped by 67.73 percent when the number of single-phase smart meter clusters doubled. The outcomes presented in Table 3 demonstrate the ability of this approach to expand the proposed solution to networks with a large number of loads. Note that it is quite burdensome to solve this problem for such a substantial number of nodes without using this approach. Moreover, by grouping the smart meters, this approach also compensates for the limited smart meter penetration percentage in networks that are not yet fully equipped with AMI, or to identify the phase of other unmetered loads such as streetlights. The only data necessary for generating the clusters is the geographical coordinates, the type of line, and the nearby loads, which in this case are generated synthetically. However, the proposed model does not depend on the topology information.

Table 3. The impact of clustering 2035 smart meters on the performance of the model

N	T	Meters Clusters	Computation Time (s)	Accuracy
2035	640	218	274.388	100%
	960	109	850.273	100%

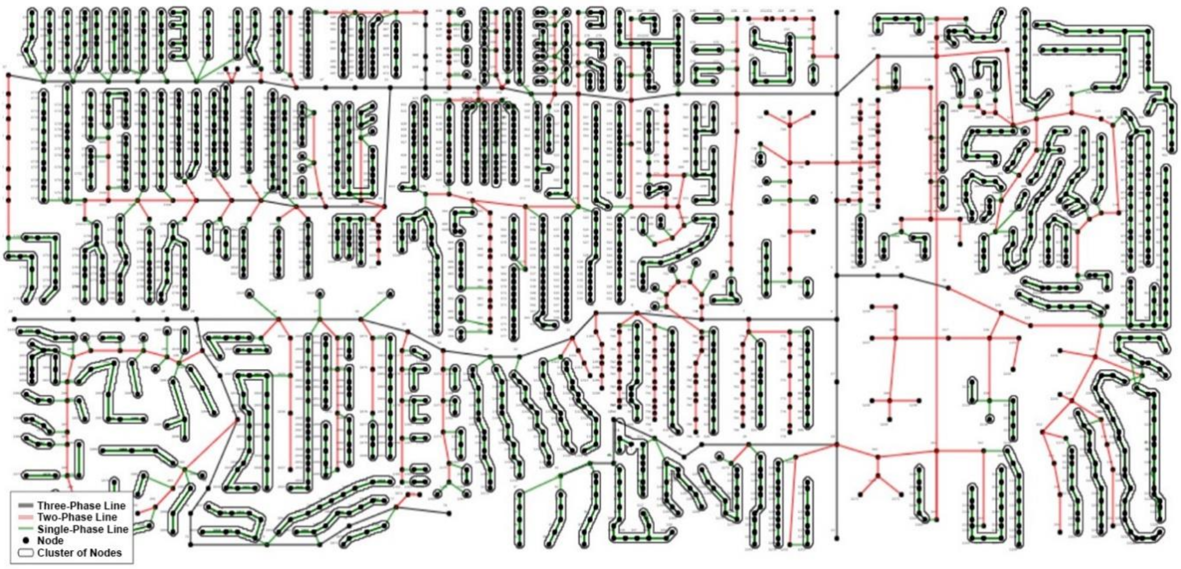

Figure 4. Topology diagram illustrating the allocation of 218 groups of single-phase nodes in a 2500-Bus network

4. CONCLUSION

The purpose of this paper was to develop a data-driven model that can identify the network topology for a realistic scenario involving a large number of consumers. The studies showed that it becomes challenging to solve the proposed optimization problem as the number of variables increases. The optimal number of measurements based on the size of the network was estimated using a sensitivity analysis, and the effect of parameters such as the gap tolerance relaxation on the performance of the model was also explored. To gain a more realistic perspective on the analysis, distribution loss and the effect of modeling this variable with a machine learning regressor algorithm were examined. Since the objective function is sensitive to the numerical values of the measurements, the ability to incorporate distribution loss into the model is one of the most important factors in improving the accuracy. The analysis led us to conclude that discarding data points with zero values while widening the measuring window boosts the computation time and accuracy. To enhance the model’s robustness, the similarity of the values in each datapoint was also considered. Using a proposed similarity score, we selected the finest data points to solve the problem. This strategy ensured that the solver would have more distinct values to work with and would not incorrectly classify smart meters with similar values into the wrong phases. Through our proposed model, the connectivity of 1,000 consumers was efficiently identified by using 2,400 measurements (equivalent to 50 days) within only a few minutes. Additional clustering, based on consumers' geographical coordinates, made it possible to identify the network topology accurately for a network with 2,500 nodes.

BIBLIOGRAPHY

[1] A. Madureira, J. Pecas Lopes, A. Carrapatoso and N. Silva, "The new role of substations in distribution network management," *CIRED 2009 - The 20th International Conference and Exhibition on Electricity Distribution.*

[2] B. Liu, K. Meng, Z. Y. Dong, P. K. C. Wong and X. Li, "Load Balancing in Low-Voltage Distribution Network via Phase Reconfiguration: An Efficient Sensitivity-Based Approach," *IEEE Transactions on Power Delivery,* pp. 2174-2185, 2021.

[3] L. Marrón, X. Osorio, A. Llano, A. Arzuaga and A. Sendin, "Low voltage feeder identification for smart grids with standard narrowband PLC smart meters," in *2013 IEEE 17th International Symposium on Power Line Communications and Its Applications*, 2013.

[4] A. Sendin, A. Llano, A. Arzuaga and I. Berganza, "Strategies for PLC signal injection in electricity distribution grid transformers," in *Power Line Communications and Its Applications (ISPLC), IEEE International Symposium*, 2011.

[5] "Boston Spa Energy Efficiency Trial - Smart Meter Voltage Measurement Performance," Boston Spa, December 2020. [Online]. Available: https://www.northernpowergrid.com/beet.

[6] "Smart Grid Legislative and Regulatory Proceedings," U.S. Energy Information Administration, November 2011. [Online]. Available: https://www.eia.gov/analysis/studies/electricity/pdf/sreg_policies.pdf.

[7] D. Alahakoon and X. Yu, "Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey," *IEEE Transactions on Industrial Informatics,* vol. 12, no. 1, pp. 425-436, 2016.

[8] A. B. Pengwah, L. Fang, R. Razzaghi and L. L. H. Andrew, "Topology Identification of Radial Distribution Networks Using Smart Meter Data," *IEEE Systems Journal,* pp. 1-12, 2021.

[9] S. Bolognani, N. Bof, D. Michelotti, R. Muraro and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *52nd IEEE Conference on Decision and Control*, 2013.

[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.

[11] Z. S. Hosseini, A. Khodaei and A. Paaso, "Machine Learning-Enabled Distribution Network Phase Identification," *IEEE Transactions on Power Systems,* vol. 36, no. 2, pp. 842-850, 2021.

[12] S. J. Pappu, N. Bhatt, R. Pasumarthy and A. Rajeswaran, "Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data," *IEEE Transactions on Smart Grid,* vol. 9, no. 5, pp. 5113-5122, 2018.

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

[14] A. Hoogsteyn, M. Vanin, A. Koirala and D. Van Hertem, "Low Voltage Customer Phase Identification Methods Based on Smart Meter Data," 2022.

[15] S. P. Jayadev, A. Rajeswaran, N. P. Bhatt and R. Pasumarthy, "A Novel Approach for Phase Identification in Smart Grids Using Graph Theory and Principal Component Analysis," 2016.

[16] M. Vanin, T. Van Acker, R. D'hulst and D. Van Hertem, "Phase Identification of Distribution System Users Through a MILP Extension of State Estimation," 2022.

[17] "How much electricity is lost in electricity transmission and distribution in the United States?," U.S. Energy Information Administration, [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=10.

[18] "SmartMeter Energy Consumption Data in London Households," UK Power Networks, [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households. [Accessed 2022].