Data-Driven Topology Generation with Physics-Guidance in LV Distribution Networks

Dong Liu[1], Juan S. Giraldo[2], Peter Palensky[1], Pedro P. Vergara[1]
[1]Intelligent Electrical Power Grids, Delft University of Technology, The Netherlands
[2]Energy Transition Studies, Netherlands Organisation for Applied Scientific Research, The Netherlands
Emails: D.Liu-7@tudelft.nl, Juan.Giraldo@tno.nl, P.Palensky@tudelft.nl, P.P.VergaraBarrios@tudelft.nl

***Abstract*—Low-voltage distribution networks (LVDNs) topology is significant for distributed energy resources (DERs) integration, and network operation management, among others. However, topology identification is a difficult task due to the outdated recordings of networks, the uncertainty of DERs and data privacy. To address this issue, a data-driven topology generation approach is proposed based on open GIS and voltage magnitude data. The proposed approach aims to generate a topology with an accurate number of main feeders and sub-branches for adjacent substations. The boundaries between adjacent substations are first identified by using hierarchical clustering (HC) to cluster normalized voltage magnitude. Given the boundaries and the location of LV transformers, a hierarchical minimum spanning tree algorithm (HMST) is adopted to generate graph topologies using GIS data, which simultaneously verifies the number of cables under the streets. Finally, the endpoints of each feeder are estimated by clustering the transformed Pearson correlation coefficient of voltage magnitude. The feasibility of the proposed approach is evaluated on two real LVDNs in the Netherlands.**

***Index Terms*—low-voltage distribution network, topology generation, correlation analysis, hierarchical clustering**

I. Introduction

Low-voltage distribution network topology is fundamental for operation management and control, such as hosting capacity analysis of DERs, congestion management, etc. However, the topology of DNs is not always available due to missed and outdated recordings. The assumptions in topology identification methods in MV networks and transmission networks make them not suitable for LVDNs [1], such as straight connection lines between transformers and availability of a large amount of measurements. Smart meter (SM) data are limited in LVDNs due to the low deployment ratio of SM and data privacy. Moreover, the increasing number of DERs leads to bi-direction power flow, challenging the identification of LVDN topology [2]. Thus, flexible topology identification methods are required to reveal the topology of LVDNs.

Data-driven topology identification approaches relying on GIS and SM data are proposed to address this issue. Open GIS data provides accessible data to identify the deployment of cables. The outline of streets is assumed to be the potential deployment ways for underground cables [3]–[5]. An optimization model was proposed to recognize the connection lines based on Open Street Map (OSM) data in [3]. To extract multiple voltage level networks using OSM data, a comprehensive data-driven method was introduced [4]. Based on detailed GIS data in specific countries, benchmark networks were generated [5], while the application of these approaches is subjected to the detailed GIS data, and the generated networks lack representation of networks in other countries. Besides, the topology extracted from OSM data only reveals the connection among buildings. The number of feeders and their sub-branches is assumed to be the same as the number of streets, which is not always true. Moreover, the boundary of substations is assumed to be known in the above approaches, while it may not be explicit in the GIS database.

Given Micro-phasor measurement units and SM data, a topology identification approach based on an alternating direction method of multipliers is proposed in [6] to jointly estimate topology and the network's parameters. A regression-based topology identification approach in [7] identifies the connection information and line impedance by recognizing the non-zero elements in the impedance matrix from SM data. Nevertheless, the above approaches assumed that a complete time-series SM dataset (i.e., voltage magnitude, active power, and reactive power) is available, which is an unrealistic assumption. Moreover, since the correlation of voltage magnitude from the same substation is stronger than that from different substations [8], correlation analysis is normally used to distinguish the voltage magnitude profiles from different regions. However, the weak correlation among the voltage magnitudes from adjacent substations decreases the accuracy of the clustering approaches that rely on Pearson correlation coefficients (PCC). Furthermore, the feeder identification issue is not considered in the aforementioned papers due to the assumption that one main feeder connects to the transformer.

To fill this gap, a data-driven topology generation with a physics-guidance approach is introduced in this paper. The proposed approach consists of three steps: boundary identification, graph topology generation and feeder identification. In the first step, the users located in adjacent substations are distinguished by clustering normalized voltage magnitude, and the boundary is determined according to the coordinates of buildings. Given the boundaries and the location of transformers, a graph topology with an accurate number of sub-branches is generated by a simplified HMST algorithm. Then, the endpoints of feeders are recognized by hierarchically clustering the PCC of raw voltage magnitude measurements. Finally, the proposed approach is tested on two real LVDNs in the Netherlands.

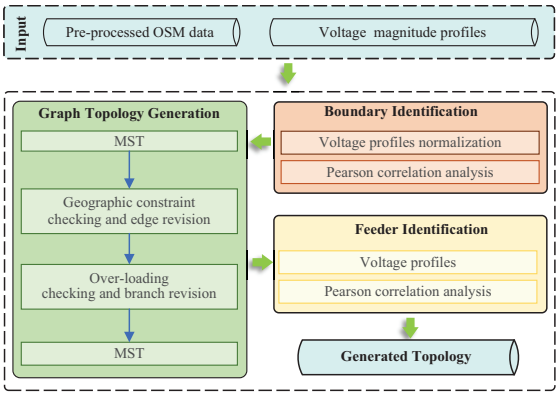

Fig. 1. Framework of proposed topology identification approach.

II. Topology Generation Framework

The proposed topology identification approach is composed of three steps: boundary identification, graph topology generation, and feeder identification. As illustrated in Fig.1.

A. Substation Boundary Identification

The proposed boundary identification algorithm integrates the HC algorithm and the correlation analysis based on the modified Pearson correlation coefficient (MPCC). The voltage magnitude is stored in a matrix V, as shown in Eq. (1). Vector $V_n = [v_{n,1}, v_{n,2}, ..., v_{n,\mathcal{T}}]^T$ represents the time series voltage magnitude at household n. $v_{n,t}$ represents the voltage magnitude at household n at time t.

$$V = [V_1, V_2, ..., V_N]^T \tag{1}$$

To mitigate the impact of voltage magnitude on the PCC values and clustering, voltage magnitude is normalized using Z-score normalization. The matrix V^* represents the normalized data and is the input for the MPCC-based HC algorithm. The row V_n^* represents a sample in the HC algorithm, and the number of clusters is set to the same as the number of substations. The correlation $PCC(V_n^*, V_m^*)$ between voltage magnitudes V_n^* and V_m^* is obtained by Eq. (2).

$$PCC(V_n^*, V_m^*) = \frac{Cov(V_n^*, V_m^*)}{\sigma_n \sigma_m} \tag{2}$$

where σ_n and σ_m are the standard deviations of samples V_n^* and V_m^*, respectively. $Cov(\cdot)$ is the covariance function.

To amplify the difference between voltage from different substations, a non-linear distance D is introduced to replace the Euclidean distance in the traditional HC algorithm.

$$D(V_n^*, V_m^*) = 1 - min\{\frac{1}{4}\ln(1 + e^{a+4\cdot PCC(V_n^*, V_m^*)}), 1\} \tag{3}$$

The second item in Eq. (3) is a modified likelihood function $F(\cdot)$ [9], which is used to calculate the MPCC of samples. Since the range of function $F(\cdot)$ is $(0, 1]$, the range of distance $D(\cdot)$ is also $[0, 1)$. The closer $D(V_n^*, V_m^*)$ is to 0, the more likely it is that the two samples V_n and V_m are collected from the same substation. Besides, the parameter a and its impact will be discussed in Section III.

Algorithm 1: MPCC-based Hierarchical clustering

Input: V^*, k_s, N
N_c = N
for $N_c \geq 1$ **do**
  for $n \leq N_c$ **do**
    for $m \leq N_c$ **do**
      $D_0(n,m)$ = max $\{D(V_i^*, V_j^*)$: $V_i^* \in \mathscr{C}_n$ and $V_j^* \in \mathscr{C}_m\}$
    end
  end
  n^*, m^*, $D_0^* \leftarrow \min\{D_0\}$
  $\mathscr{C}_n^* \leftarrow \mathscr{C}_n \cup \mathscr{C}_m$
  N_c = N_c -1
  $L_m[N - N_c] \leftarrow (n^*, m^*, D_0^*, N^*)$
end
$\mathscr{L} \leftarrow fcluster(L_m, k_s)$
Output: Cluster: C_1, ... ,C_{k_s}

The MPCC-based HC algorithm is shown in Algorithm 1. The calculation in the outer loop is to obtain the linkage matrix of the input data. The final line is to cluster the input data into a k_s cluster and obtain the labels $\mathscr{L}$ by the traditional HC algorithm (i.e., the function *fcluster* in Scip).

B. Graph Topology Generation

Two common deployment styles of cables in LVDNs are depicted in Fig. 2 [10]. A single cable is deployed under streets with households on only one side in Fig. 2 (a), while two cables are deployed under streets with households located on both sides in Fig. 2 (b). To generate a radial topology with an accurate number of sub-branches, the HMST algorithm proposed in our previous work [11] is adopted.

The input of the HMST algorithm consists of the shortest path matrix P_{LV} between households and transformers and the shortest path matrix P among households. The matrix P is used as the weight of edges while constructing the graph topology. The peak demand-based refinement strategy in HMST aims to verify the number of cables under streets based on the maximum capacity of cables and peak demand. The maximum load I_s of the street s is estimated using the expression in (4).

$$I_s = \frac{(r_g)^k \cdot N_s \cdot C_o \cdot P_{pe}}{3 \cdot cos\theta \cdot V_0} \tag{4}$$

where r_g is the annual growth of demand and k is the planning period. C_o represents the concurrency for N_s houses, representing how many households reach peak load simultaneously. P_{pe} is the average peak demand. $cos\theta$ is the power factor and V_0 is the voltage level of the street.

In the first step, the traditional MST algorithm is used to generate a radial tree with the shortest length of cables, represented by $MST(\cdot)$ and the edge in the tree is represented by T_w. The weight in P is then adjusted to ensure that all edges in the generated tree are in P_{LV}. In the second step, if I_s is larger than the maximum capacity $\bar{I}$ of the deployed cable, two cables are assigned for this street, and two sub-trees T_0 for

(a) (b)

Fig. 2. Two deployment styles of cables in LVDNs.

Algorithm 2: FPCC-based Hierarchical Clustering

Input: V, k_f, N
for $n \leq N$ **do**
  for $m \leq N$ **do**
    $\rho^*_{n,m}$ = FPCC(V_n,V_m)
  end
end
$\mathscr{L} \leftarrow$ **Algorithm1**(P^*,k_f,N)
Output: Cluster: C_1, ... ,C_{k_f}

the households located on each side are obtained. Conversely, one tree is generated for all households. The topology T of the main feeders is obtained based on the updated matrix P. The graph topology is obtained by combining the T_0 and T.

C. Feeder Identification

The endpoints of feeders cannot be directly inferred from the OSM data since the length and the deployment of each feeder are not recorded or missed. Compared to voltage magnitudes from different feeders, the PCC shows a higher correlation value among voltage magnitudes in the same feeders [8]. This characteristic means that the voltage magnitude from the same feeder shows similar correlations. Inspired by this, an FPCC-based HC algorithm that integrates Fisher z-transformation and PCC analysis is proposed, as shown in Algorithm 2. To amplify the difference between voltage correlation from different feeders, a modified Fisher z-transformation function is employed to transform PCC, as shown in line 3 in Algorithm 2, which is formulated as:

$$FPCC(V_n,V_m) = \ln\left(\frac{1+PCC(V_n,V_m)}{1-PCC(V_n,V_m)+\alpha}\right) \tag{5}$$

where α added to the denominator is used to avoid an infinite value of FPCC and to control its distribution region.

The number of feeders connected to the transformer is assumed to be known and taken as the number of clusters. The input of Algorithm 2 is the voltage magnitude rather than the normalized voltage magnitude. The input of the integrated Algorithm 1, as shown from line 6 in Algorithm 2, is the transformed PCC matrix. Each row P_n in P replaces the corresponding row in V in Algorithm 1. The output of Algorithm 2 is the k_f clusters and the households in each cluster that are located farthest from the LV transformer are the endpoints of each feeder.

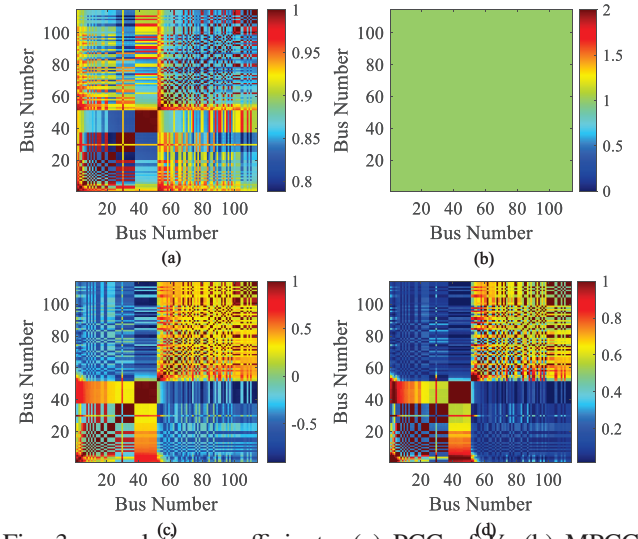

Fig. 3. correlation coefficients: (a) PCC of V, (b) MPCC of V, (c) PCC of V^* and (d) MPCC of V^*.

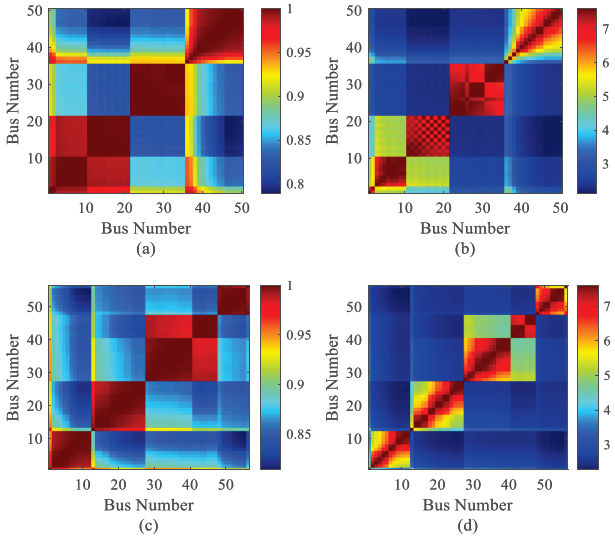

Fig. 4. PCC and FPCC in two LVDNs: (a) PCC in LV-52, (b) FPCC in LV-52, (c) PCC in LV-62 and (d) FPCC in LV-62.

III. Case study

The proposed approach is evaluated on two adjacent LVDNs in the Netherlands [10]. The two LVDNs consist of 52 and 62 connection points and are named LV-52 and LV-62, respectively. The load profiles with a resolution of 15 minutes are selected and scaled from reference [12], and the $cos\theta$ is set at 0.95. The voltage magnitude profiles are generated by solving a power flow model [13]. The proposed approach is implemented in Python. The linkage criteria in the traditional HC algorithm are set as complete linkage. The parameter a is set as the same value in [9], and parameter α is set as 0.01.

A. Correlation Evaluation

The PCC and MPPC of unnormalized and normalized voltage magnitudes are shown in Fig. 3. Compared to the PCC of unnormalized voltage magnitude in Fig. 3(a), there is a clear boundary in the distribution of PCC of normalized

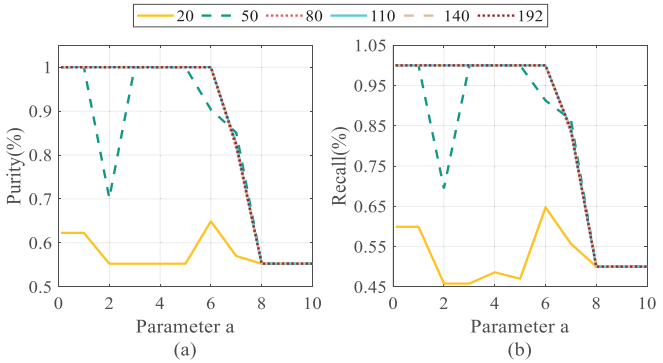

Fig. 5. Purity and Recall of A1: (a) Purity and (b) Recall.

voltage magnitude in Fig. 3(c), which indicates the boundary of substations. Scaled by Eq. (3), the PCC in Fig. 3(a) becomes 1. However, the MPCC of normalized voltage magnitudes in Fig. 3(d) shows a more clear boundary. As shown in Fig. 3(b) and (d), Z-score normalization amplifies the difference between voltage magnitude profiles from different substations while attenuating the difference between voltage magnitude profiles from different feeders. Thus, unnormalized voltage magnitude data are more suitable for feeder identification.

The PCC and the FPCC of unnormalized voltage magnitude are depicted in Fig. 4. Although there are clear boundaries in Fig. 4(a) and (c), the samples around the boundaries may be misidentified due to the higher correlation with the start points of the other feeders, such as the 1st household and the 40th household in Fig. 4(a).The FPCC in LV-52 and LV-64 in Fig. 4(b) and (d) also show more clear boundaries. Moreover, there is a significant disparity in the FPCC values on either side of the boundary. Additionally, compared to the raw voltage magnitudes, the FPCC has a better representation of the unique characteristics of feeders. For instance, for the first feeder in LV-62, the FPCC vector P_n^* (n=1,...,10) exhibits higher magnitudes in dimensions 1 to 10, with relatively smaller values in other dimensions.

B. Performance of Proposed Approach

The goal of the proposed approach is to generate topology by following a three-step approach. The parameter a in Eq. (5) impacts the calculation of the linkage matrix L_m and further impacts the accuracy of boundary identification. As common indicators for evaluating clustering algorithms, the purity P_{pu} and average recall R of Algorithm 1 are calculated to analyze the impact of parameter a.

The curves of purity and average recall in Fig. 5 decrease with the increasing of parameter a. When the purity P_{pu} or average recall R is around 0.5, it means that the input voltage magnitude from the two substations is classified into the same clusters, i.e., the proposed Algorithm 1 fails to identify the boundary of the substations. In particular, when 20-dimension voltage magnitude vectors V_n are available, Algorithm 1 fails to identify the boundaries. On the other hand, when parameter a is set between 5 and 8, the purity and average recall

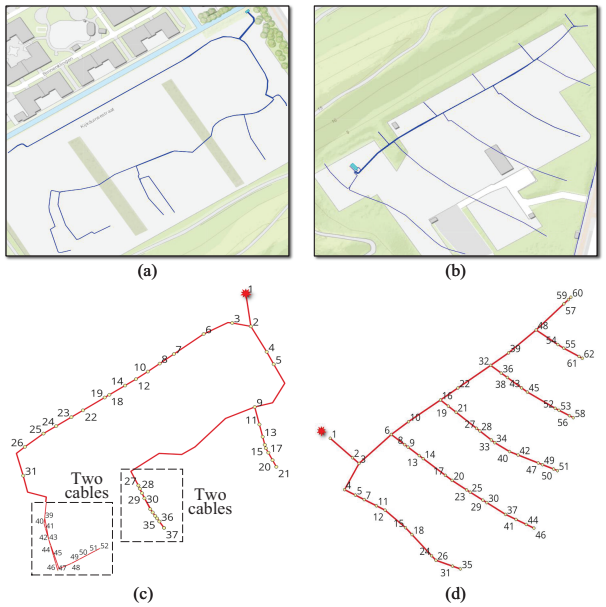

Fig. 6. Topology for (a) actual topology for LV-52, (c) generated topology for LV-52, (b) actual topology for LV-62 and (d) generated topology for LV-62.

TABLE I: The Purity of A3 With Different α.

Parameter	Dimension of Voltage					
α	20	50	80	110	140	192
0.0001	1.00	0.98	0.98	0.98	0.98	0.98
0.0010	**1.00**	**0.98**	**1.00**	**1.00**	**1.00**	**1.00**
0.0100	1.00	1.00	1.00	0.98	0.98	0.98
0.1000	1.00	1.00	1.00	0.98	0.98	0.98

significantly decrease until around 0.5. Thus, parameter a should be set to a value less than 5.

Based on the obtained boundaries, the generated graph topologies for LV-52 and LV-62 are shown in Fig. 6. As shown in Fig. 6(c), two cables under streets 2 and 3 are identified. However, compared to the actual topology, there is an inaccurate connection line at the 39th node in the generated topology of LV-52, which is caused by deployment-related factors and other physical constraints. The path between the 39th node and the 31st node is shorter than the path between the 39th node and the 27th node. However, in the actual network, the 39th node is connected to the 27th node. The generated topology for LV-62 in Fig. 6(d) is consistent with the actual topology in Fig. 6(b).

There are four feeders in LV-52 and LV-62, respectively. To analyze the impact of parameter α and voltage magnitude dimension on the Algorithm 2, the purity of Algorithm 2 is analyzed under multiple scenarios. Table I presents the purity under voltage magnitude profiles from LV-52. The minimum purity of Algorithm 2 is 98%, and the purity remains at 100% as more data becomes available or as parameter α is tuned. In particular, when parameter α is set to 0.0010, Algorithm 2 identifies each feeder using more than one day's voltage magnitude. Meanwhile, the 39th node is clustered in the same cluster as the 27th node instead of in the same cluster as the

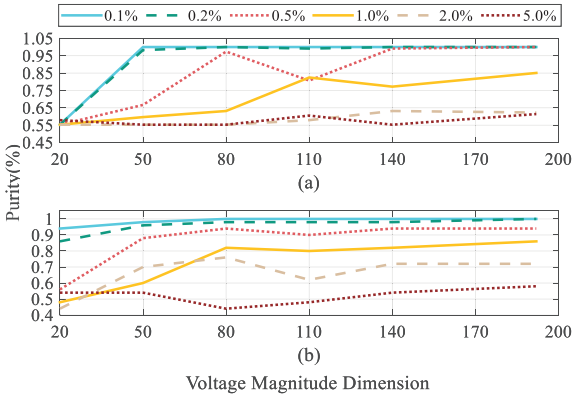

Fig. 7. Purity of A1 and A3 under multiple SM classes: (a) Purity of A1 and (b) Purity of A3 in LV-52.

31st node, which means the inaccurate connection lines in the generated graph topology are revised. Given the voltage magnitude in LV-62, the purity remains at 100% independent of parameter α, except for the 50-dimension voltage magnitude. In summary, Algorithm 2 efficiently identifies each feeder in DNs when the parameter α is set around 0.0010 and the dimension of the input voltage data is greater than 96.

C. Robustness Analysis

Considering the random error brought by SMs, six kinds of Gaussian error (i.e., $v_e \sim \mathcal{N}(\mu, \sigma)$) are generated and added to the simulation voltage magnitude data according to the accuracy requirements for SMs [9]. The mean μ of the Gaussian distribution is set as 0. According to the 3σ principle, three times the standard deviation 3σ of the Gaussian distribution is set as 0.2%, 0.5%, 1%, 2% and 5%.

Fig. 7 presents the purity of Algorithm 1 and Algorithm 2 under voltage magnitude data with the above errors. Given data from the high-precise SMs (e.g., 0.2%, 0.5%), Algorithm 1 can identify the accurate boundaries of substations, while Algorithm 1 fails to identify the boundaries and feeders using data from the low-precise SMs (i.e., 1%, 2% and 5%). According to Fig. 7(b), Algorithm 2 is more sensitive to error compared to Algorithm 1. Based on voltage magnitude with smaller magnitude errors, Algorithm 2 can identify at least 85% of the user-to-feeder relationship using more than two days' data. However, Algorithm 2 only identify around 50% user-to-feeder relationship using voltage magnitudes from the low-precise SMs (i.e., 1%, 2% and 5%). Thus, the two proposed algorithms are more suitable for DNs with high-precise SMs. Besides, the collected voltage magnitude data may be incomplete due to the communication issues in the cyber layer. The purity of Algorithm 1 and Algorithm 2 under incomplete voltage magnitude are summarized in Table II. The impact of missing data is similar to the impact of the voltage dimension. Given more than one day's data, the accuracy of the proposed algorithms remains at 100%.

TABLE II: Purity Under Incomplete Voltage Magnitude data.

Algorithm	Voltage Dimension	Incomplete rate (%) 5	10	20	30	40	50
A1	50	0.99	0.99	0.99	0.95	0.9	0.97
	100	1.00	1.00	1.00	1.00	1.00	1.00
A3(LV-52)	50	0.99	0.98	0.99	0.98	0.98	0.98
	100	1.00	1.00	1.00	1.00	1.00	1.00
A3(LV-62)	50	0.98	0.98	0.99	0.99	0.99	0.98
	100	1.00	1.00	1.00	1.00	1.00	0.99

IV. Conclusion

A topology identification approach is proposed based on voltage magnitude and GIS data to extract topologies that are close to the actual topologies in LVDNs. The boundaries of substations are first identified by correlation analysis of voltage magnitude, and the graph topology is generated based on OSM data and the obtained boundaries. The inaccurate connections in the generated graph topologies induced by the mesh streets are revised by correlation analysis simultaneously in the feeder identification step. With the guidance of GIS data, the generated topology not only presents the connections but also reveals the deployment of the cables in LVDNs. The results show that the generated topologies using OSM and incomplete voltage magnitudes approximate the actual topology. Our future work aims to identify topology with parallel feeders and meshed structures.

References

[1] H. Zhang, J. Zhao, X. Wang, and Y. Xuan, "Low-voltage distribution grid topology identification with latent tree model," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 2158–2169, 2022.

[2] G. Cavraro and R. Arghandeh, "Power distribution network topology detection with time-series signature verification method," *IEEE Transactions Power Systems*, vol. 33, no. 4, pp. 3500–3509, 2017.

[3] H. K. Çakmak, L. Janecke, M. Weber, and V. Hagenmeyer, "An optimization-based approach for automated generation of residential low-voltage grid models using open data and open source software," in *2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2022, pp. 1–6.

[4] J. Kays, A. Seack, T. Smirek, F. Westkamp, and C. Rehtanz, "The generation of distribution grid models on the basis of public available data," *IEEE Transactions Power Systems*, vol. 32, no. 3, pp. 2346–2353, 2017.

[5] G. Pisano, N. Chowdhury, M. Coppo, N. Natale, and F. Pilo, "Synthetic models of distribution networks based on open data and georeferenced information," *Energies*, vol. 12, no. 23, p. 4500, 2019.

[6] P. Shah and X. Zhao, "Network identification using μ-PMU and smart meter measurements," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7572–7586, 2022.

[7] X. Wang, Y. Zhao, and Y. Zhou, "A data-driven topology and parameter joint estimation method in non-pmu distribution networks," *IEEE Transactions Power Systems*, 2023.

[8] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964–1971, 2015.

[9] S. García, J. M. Mora-Merchán, D. F. Larios, E. Personal, A. Parejo, and C. León, "Phase topology identification in low-voltage distribution networks: A bayesian approach," *International Journal of Electrical Power & Energy Systems*, vol. 144, p. 108525, 2023.

[10] Stedin. Accessed: Aug. 2023. [Online]. Available: https://www.stedin.net/zakelijk/open-data/liggingsdata-kabels-en-leidingen

[11] D. Liu, J. S. Giraldo, P. Palensky, and P. P. Vergara, "Topology identification and parameters estimation of lv distribution networks using open gis data," *Available at SSRN 4661179*, 2023.

[12] K. P. Schneider, B. Mather, B. C. Pal, C.-W. Ten, G. J. Shirek, H. Zhu, J. C. Fuller, J. L. R. Pereira, L. F. Ochoa, L. R. de Araujo *et al.*, "Analytic considerations and design basis for the IEEE distribution test feeders," *IEEE Transactions Power Systems*, vol. 33, no. 3, pp. 3181–3188, 2017.

[13] P. P. Vergara, J. C. López, M. J. Rider, and L. C. Da Silva, "Optimal operation of unbalanced three-phase islanded droop-based microgrids," *IEEE Transactions Smart Grid*, vol. 10, no. 1, pp. 928–940, 2017.