Topology identification and parameters estimation of LV distribution networks using open GIS data

Liu, Dong; Giraldo, Juan S.; Palensky, Peter; Vergara, Pedro P.

DOI
10.1016/j.ijepes.2024.110395

Publication date
2025

Document Version
Final published version

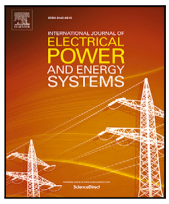
Published in
International Journal of Electrical Power and Energy Systems

Electrical Power and Energy Systems 164 (2025) 110395

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes



Topology identification and parameters estimation of LV distribution networks using open GIS data

Dong Liu [a], Juan S. Giraldo [b], Peter Palensky [a], Pedro P. Vergara [a,*]

[a] Intelligent Electrical Power Grids (IEPG) Group, Delft University of Technology, 2628CD, The Netherlands
[b] Energy Transition Studies Group, Netherlands Organisation for Applied Scientific Research, 2595 DA, The Netherlands

ARTICLE INFO

Keywords:
Distribution networks
Topology generation
Open source data
Incomplete data
Optimization

ABSTRACT

The topology of low-voltage distribution networks (LVDNs) is crucial for system analysis, e.g., distributed energy resources (DERs) integration, network hosting capacity analysis, state estimation, and electric vehicle charging management. However, it is frequently unavailable or incomplete. This paper develops a data-driven topology identification approach for LVDNs with a high proportion of underground cables. The proposed approach exploits the fact that underground cables usually follow the street pattern, thus relying on open street map (OSM) and smart meter (SM) data. Three stages compose the proposed approach: In the first stage, a hierarchical minimum spanning tree algorithm is proposed to generate the initial topology with an accurate number of sub-branches from the pre-processed OSM data and peak demand. In the second stage, based on the limited SM data, the location of breakpoints in mesh topology caused by circle roads is verified and reconstructed to guarantee the radial structure of LVDNs. Finally, given multiple incomplete SM datasets, three data-driven optimization models based on a state estimation model are constructed to mitigate the error of cable length induced by OSM data. The feasibility of the proposed topology identification approach is verified on three actual LVDNs in The Netherlands and multiple incomplete SM datasets. Furthermore, the minimal amount of SM data needed to minimize the error of cable length is analyzed.

1. Introduction

Distribution network topology is fundamental for distribution system operators (DSOs) in operation analysis, DERs hosting capacity analysis and integration, among other applications. However, DSOs usually do not keep full records of the updated topology due to wrong, missed, or outdated recordings. On top of that, the increasing uncertainty of DERs, including household photovoltaic (PV) systems, electric vehicles (EVs), etc. [1,2], impacts the topology reconfiguration frequency and the relationship between measurements, challenging the identification of LVDN topology. Moreover, the low deployment rate of SM hinders the application of the topology identification methods used in transmission networks and MV networks [3,4], usually developed considering assumptions not suitable for LVDNs, such as available connection points (i.e., the location of the MV/LV transformer), straight connection lines between transformers, etc. Although open synthetic networks and benchmark models are available [5,6], flexible topology identification methods for already deployed LVDNs are required.

To overcome this challenge, multiple topology identification methods are proposed, which are roughly classified into SM data-based and open data-based methods. In SM data-based methods [7–14], topology identification is considered a binary classification problem or regression problem that aims to identify the stage of the edge and switches. The complete SM data of each user is always assumed to be known and accessible, including the time-series voltage, active power, and a part of the connecting information of the networks. Meanwhile, the measurements are used as synchronous data. Based on a state estimation model and a regression model, a hybrid topology identification approach was proposed to handle the impact of the uncertainty of DERs [7]. To restrict the propagation of SM data error, a probabilistic graphical model-based topology estimation method was developed in [8]. However, these methods only reveal the connection between buses without estimating line parameters, which is a necessary part of modeling an LVDN using a power flow formulation.

Based on an alternating direction method of multipliers, the work in [9] proposed a robust topology identification method to jointly estimate the network's parameters and its topology, relying on μ-PMUs and SM datasets. While voltage angles are normally unavailable in the distribution network (DN), [10] establishes a data-driven model based on an impedance matrix to identify topology and regress the line parameter from SM data. In [11], a hybrid data-driven approach

* Corresponding author.
E-mail address: P.P.VergaraBarrios@tudelft.nl (P.P. Vergara).

https://doi.org/10.1016/j.ijepes.2024.110395
Received 27 November 2023; Received in revised form 17 September 2024; Accepted 19 November 2024
Available online 2 December 2024

integrating a partial correlation analysis strategy and a linear regression model is proposed to generate topology from limited SM data. Considering the error in SM data and dynamic topology changes while collecting measurements, a maximum-likelihood-based joint estimation approach is established [12]. Based on multiple linear regression models, [13] presents a comprehensive topology identification approach to simultaneously estimate the topology, line parameters, and phase connection from raw SM data. Although the above methods can generate an accurate topology, the requirements of SM data and prior topology information (e.g., the topology candidates) make them infeasible in practice due to the low deployment rate of SM in LVDNs and data privacy-related problems. Moreover, the generated topology, without using geographic information systems (GIS) data, may not accurately depict the actual deployment of connection lines. Furthermore, the resistance and reactance of the lines are optimized as independent variables, which is not suitable for accurate estimation of the deployed cable. In general, the impedance ratio of the deployed cable is fixed and may be slightly influenced by environmental factors, such as temperature.

In the second category, LVDN topologies are generated making use of open GIS data and planning rules [15], such as Open Street Maps (OSM) [16], OpenGridMap [17], etc. Well-designed physical constraints, such as the cable routing based on street layouts and building locations, can significantly enhance the accuracy of the generated LVDN topologies [18]. To make the topology inferred from OSM data match the actual structure of LVDNs, a simplified optimization model for line power flow is designed and verified in [19]. A comprehensive method is proposed in [20] to generate large-scale distribution networks with different voltage levels. Besides, based on detailed GIS data, some methods are introduced to generate benchmark models or representative networks in specific countries [21–23]. However, the characteristics of LVDNs vary among different countries, so the generated topology based on supply tasks in one country may not be consistent with the LVDNs in other countries [24]. Additionally, the identification accuracy of LVDN topology is influenced by various geographic constraints specific to different regions, such as the proportion of underground cables, the number of road loops, etc. Specifically, the complex mesh roads in urban environments challenge the construction of radial topologies, i.e., the identification of breakpoint locations. Although these methods based on open GIS data show high performance without relying on SM data, the generated topology is only consistent with the topology of the initial stage of construction, which is one of the main disadvantages. This means that the extracted topology is still outdated without further modification based on the latest SM data. Moreover, the error in the cable length caused by the missed or inaccurate OSM data is not optimized in the above papers, which assume the location of connection points on the street is correct.

To fill this research gap, this paper introduces a data-driven topology identification approach that leverages the strengths of the aforementioned two kinds of approaches. The proposed approach consists of three stages: graph topology generation, topology reconfiguration, and topology optimization. In the first stage, the aim is to generate an initial topology with an accurate number of cables from OSM data that satisfies geographic constraints, e.g., no connecting lines cross a building or a river. To do this, this paper uses a hierarchical minimum spanning tree (HMST) algorithm to check the number of underground cables according to the maximum capacity of the deployed cable and the peak demand. The traditional MST is first adopted to connect power connecting points along the streets while ensuring the shortest total cable length and the radial structure of the distribution networks. Then, the number of cables under the streets with households on both sides is verified and modified. In the second stage, the street-to-street connection lines are verified and reconstructed based on a voltage magnitude residual. Finally, in the third stage, to mitigate errors in cable length induced by inaccurate OSM data, three data-driven optimization models based on a power flow model [25] are constructed based on

Nomenclature

Index/Set

s	Index of streets
I_0	Index of the non-root nodes of trees
I_L/I_H	Index of nodes with negative/positive voltage magnitude residual
I_{L1}/I_{H1}	Index of 1-degree nodes of cables with negative/positive residual
I_{L1I}/B_L	Index/set of start nodes of cables with negative residual
I_{H1I}/B_H	Index/set of start nodes of cables with positive residual
I_{NP}	Index of the nearest upstream node of the node in B_h
$i/\bar{i}$	Index of iteration and its threshold
$mn/\mathcal{L}/\mathcal{L}_s$	Index/set of mainline and service line
C	Index vector of SM
l_{mi}	Connecting line at middle of the ith branch
$l_{mi,3}$	Connecting line in the ith sub-branch between start point and middle point
$m/\mathcal{N}$	Index/set of nodes in the networks
$t/\mathcal{T}$	Index/set of time step
D^n_{mn}	Set of lines from transformer to bus n

Parameter

$N_0/N_t/N_s$	Number of houses/branches/streets
N_{un}	Number of unmetered houses
$D/D_{s,i}$	Shortest path matrix among all buildings/ buildings located at the ith side of street s
D_{LV}	Shortest path matrix between buildings and the transformer
γ_0	Flag parameter for street, i.e., 0/1
S_0	Number of streets with buildings on two sides
c_i	Flag parameter for SM, i.e., 0/1
r_g	Annual growth of demand
C_0	Concurrency factor for households
P_{pe}	Average peak demand
$cos\theta$	Power factor of households, set as 0.95.
$\overline{I}$	Maximum capacity of the deployed cable
k	Planning period
r_{mn}/x_{mn}	Real resistance/reactance of lines
$\widetilde{l}_{mn}$	Extracted cable Length from OSM data
$\widetilde{r}_{mn}/\widetilde{x}_{mn}$	Resistance/reactance of the cable whose length is extracted from OSM data
T	Dimension of time-series data
R_{unm}	Unmetered rate in DN
$\overline{n}$	Threshold for variable n^*
$\Delta\overline{V}$	Threshold for voltage magnitude residual
$P^{D0}_{m,t}/Q^{D0}_{m,t}/V^0_{m,t}$	Active/reactive power/voltage magnitude measurement at node m at time step t
$\overline{\alpha}_{mn}/\underline{\alpha}_{mn}$	Upper/lower limit for mainline length ratio
$\underline{\beta}_{mn}/\overline{\beta}_{mn}$	Upper/lower limit for service line length ratio
T^*	Pseudo-time horizon
T^d	Dimension of daily sample
$\overline{V}/\underline{V}$	Upper/lower limit for voltage magnitude
$\underline{V}^0_{m,t}/\overline{V}^0_{m,t}$	Daily minimum/maximum voltage magnitude
$\underline{P}^{D0}_{m,t}/\overline{P}^{D0}_{m,t}$	Daily minimum/maximum active power
$\underline{Q}^{D0}_{m,t}/\overline{P}^{D0}_{m,t}$	Daily minimum/maximum reactive power
$w_v/w_p/w_q$	Weight for residual of active power/reactive power/voltage magnitude
NIH	Number of households integrated in objective function

Variables	
$\Gamma/\Gamma_0/\Gamma_w$	Generated minimum spanning tree/ sub-trees/Edges of trees
G/G^*	Initial/Final Graph topology
I_s	Estimated maximum load for street s
n^*	Number of nodes with excessive residuals
$\Delta V^d_{mn,t}$	Voltage drop on line mn at time t
$\Delta V^d_{m,t}$	Total voltage drop from the transformer to bus m at time t
ΔV_m	Voltage magnitude residual at node m
ΔV^*	Maximum voltage magnitude residual
η	Margin for voltage magnitude residual
$P_{mn,t}/Q_{mn,t}$	Active/reactive power at line mn at time step t
$P^s_{m,t}/Q^s_{m,t}$	Injection active/reactive power at node m at time step t
$P^D_{m,t}/Q^D_{m,t}$	Active/reactive power at node m at time step t
$E^P_m/E^Q_m/E^V_m$	Residual of active power/reactive power/ voltage magnitude
$V_{m,t}$	Estimated voltage magnitude at node m at time step t
l_{mn}	Ratio of optimized and extracted length of cable mn
Acronyms	
HV/MV/LV	High/Medium/Low voltage
LVDN	Low-voltage distribution network
DN	Distribution network
SM	Smart meter
DSO	Distribution system operator
DERs	Distribution energy resources
GIS	Geographic information systems
PV	Photovoltaic
EV	Electrical vehicles
OSM	Open street map
HMST	Hierarchical minimum spanning tree

multiple incomplete SM datasets. The parameters used in this paper are summized in Table 1 and Table 2 summarizes the approaches discussed in the aforementioned papers and the proposed approach. The cells marked with "Y" or "N" indicate whether specific issues and data are considered in each approach. "P/Q/V/θ" represents reactive power, voltage magnitude, and phase, respectively. The proposed approach is finally tested on three LVDNs in the Netherlands and incomplete SM datasets. The main contributions of this paper are summarized as follows:

- A HMST algorithm integrating geographic constraints and a peak demand-based refinement strategy is proposed to generate a radial feasible topology from the pre-processed GIS data. The proposed refinement strategy can identify the number of cables based on the street layout, considering their maximum capacity.
- To ensure accurate radial structure, a power flow model-based topology reconstruction strategy is proposed to verify and revise street-to-street connections (i.e., the location of the breakpoints) in graph topology when there are loops in LVDNs.
- Three data-driven optimization models are proposed to mitigate the error in the length of the cables using multiple smart meters datasets, including a complete SM dataset, an SM dataset with randomly missed data, and an SM dataset composed of only daily maximum and minimum data. The minimum amount of SM data needed to identify the actual length of underground cables is analyzed.

Table 1
Parameter settings.

Parameter name	Setting approach
γ_0	Set based on OSM data, i.e., 0/1
c_i	Provided by DSO, i.e., 0/1
r_g	Set by DSO at planning period
C_0	Set by DSO or based on the statistical features of electricity usage in LVDN
P_{pe}	Average peak demand in the LVDN, provided by DSO or open websites
$\overline{I}$	Based on the type of cable, provided by DSO
k	Planning period, provided by DSO
$\widetilde{r}_{mn}/\widetilde{x}_{mn}$	Calculated based on extracted length $\widetilde{l}_{mn}$ and type of the deployed cable.
$\overline{n}$	Number of the houses located on one of the streets in LVDN, obtained from OSM data
$\Delta\overline{V}$	Estimated from historical voltage data by DSO or preliminary results
α/β	The limits of α/β are estimated based on OSM data and revised by DSO according to the available recording of cable length
$\overline{V}/\underline{V}$	Set by DSO, normally set as 0.95–1.05 p.u.
$w_v/w_p/w_q$	Obtained through cross-validation using historical data. w_v should be much larger than w_p and w_q.
R_{unm}	Set by DSO based on the number of the unmetered households.
NIH	Set by DSO based on accuracy requirement
Note:	$N_0, N_t, N_s, N_{un}, D/D_{s,i}, D_{LV}, S_0$ are extracted from OSM by Dijkstra's algorithm, and the detailed information is illustrated in Section 2.2.
	$cos\theta, T, P^{D0}_{m,t}/Q^{D0}_{m,t}/V^0_{m,t}, T^*, T^d, \underline{P}^{D0}_{m,t}/\overline{P}^{D0}_{m,t}$, $\underline{Q}^{D0}_{m,t}$ and $\overline{P}^{D0}_{m,t}$ are provided by DSO, i.e., the collected measurements.

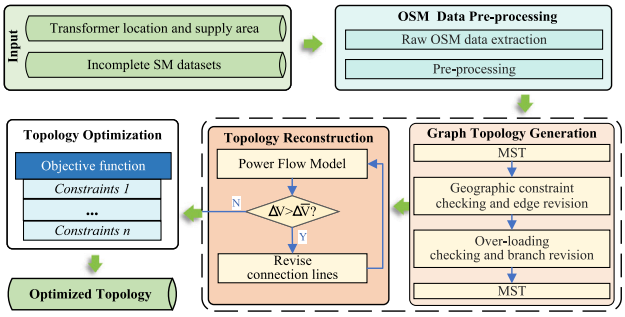


Fig. 1. Framework of the proposed data-driven topology identification approach for distribution networks: Stage I: graph topology generation, Stage II: topology reconstruction and Stage III: topology optimization.

The remainder of this paper is organized as follows: Section 2 illustrates the framework of LVDN topology identification, OSM data pre-processing, graph topology generation, topology reconstruction, and topology optimization. Besides, the reorganization of daily maximum and minimum SM data is also illustrated. Sections Section 3 describe the case of studies and results. Section 4 presents the conclusions of this paper.

2. LVDN topology identification framework

The proposed topology identification approach is illustrated in this section, which is shown in Fig. 1. The proposed approach comprises three stages: Stage I: graph topology generation, Stage II: topology reconstruction, and Stage III: topology optimization.

Given the location of the MV/LV transformer and the boundaries of the supply area, raw OSM data in the area are first extracted and pre-processed. The first stage is the application of the proposed HMST algorithm. An initial graph topology for the LVDN is generated based

Table 2
Topology identification methods comparison.

Ref.	Input			Approach				Topology	
	OSM	SMs data	Cable type	Cable outline	Sub-branch checking	Incomplete data	Length error mitigation	Voltage level	Service line
[8]	N	V	N	N	N	N	N	MV/LV	N
[9]	N	P/Q/V	N	N	N	N	Y	MV	N
[10]	N	P/Q/V	N	N	N	N	Y	MV	N
[11]	N	P/Q/V	N	Y	Y	N	Y	MV	N
[12]	N	P/Q/V/θ	N	N	N	N	Y	MV	N
[13]	N	P/Q/V	Y	Y	Y	Y	Y	LV	N
[16]	Y	N	N	Y	N	N	N	HV/MV	Y
[17]	Y	N	Y	Y	N	N	N	LV	Y
[18]	Y	N	N	Y	N	N	N	HV/MV/LV	N
[19]	Y	P	Y	Y	N	N	N	LV	Y
[20]	Y	N	Y	Y	N	N	N	HV/MV/LV	Y
Our work	Y	P/Q/V	Y	Y	Y	Y	Y	LV	Y

on pre-processed OSM data and the average peak demand. In the second stage, the locations of the breakpoints are verified and revised based on the voltage magnitude residual. The revised topology approximately reveals the connection information among the households in the LVDN, i.e., the potential outlines of underground cables. Finally, in the third stage, considering the data privacy and the existing unmetered customers, the error in cable length is mitigated by the constructed data-driven optimization models depending on the available SM data. The output of the proposed approach is a refined LVDN topology of the area, which is generated only relying on the given location of the transformer, its supply area, the type of cables, and the available SM data. The generated topology is similar to the available feeder test cases and close to the latest topology for the LVDN, which includes the connecting points, the impedance of each cable segment, and the root node index (i.e., the location of the LV transformer). Before introducing each stage in detail, the pre-processing applied to the OSM is described.

2.1. OSM data pre-processing

The process of OSM data extraction and pre-processing is depicted in Fig. 2. Given the location of the transformer and supply areas in Fig. 2(a), all related buildings and streets are first extracted as shown in Fig. 2(b). The raw OSM dataset includes the buildings and streets located outside the area, which is due to their interconnections with the streets within the supply area. These redundant elements are removed. Then, the shortest lines between streets and buildings are extracted and taken as service lines, as depicted in Fig. 2(c). The endpoints of service lines are defined as the connection points of buildings. The connection points of the buildings that are aligned in a linear arrangement should be connected to linear cables that are similarly deployed, such as the buildings located at street s in Fig. 2(c). Thus, the connection points are verified and revised. Besides, the node pairs whose distance is less than a certain threshold are merged into one node, and extra connection nodes are added to the crosspoints of streets. Finally, the basic datasets for generating a graph topology are obtained, including the outline of streets, the coordinates of buildings' centers, the connecting point of buildings, and the number of households in each building (i.e., N_0 buildings).

2.2. Stage I: Graph topology generation

Radial LVDNs with underground cables can be represented as undirected graphs, where the nodes depict the connection points of each household and the edges depict the underground cables. The outline of streets is assumed to be the potential deployment outline of underground cables, which is normally correct and verified in [18–20,26]. Dijkstra's and MST algorithms are adopted to generate a potential graph LV topology.

Dijkstra's algorithm calculates the shortest path matrix D_{LV} between buildings and the MV/LV transformer and the shortest path matrix D among buildings, which is subjected to the outline of the streets. The matrix D_{LV} is used as the geographic constraint, and the matrix D is used as the weight of edges while constructing the initial graph topology. The geographic constraints (e.g., the road segments under maintenance, the shortest path-related constraints) about the topology will contribute to the accurate LVDN topology generation. Besides, there may be two cables deployed beneath the street s with buildings on both sides. The path matrices D_s, $D_{s,1}$ and $D_{s,2}$ for the buildings located on the street s and its two sides streets are extracted from the path matrix D. The flag parameter γ_0 is set as 1. On the other hand, when there is no street with households on both sides in LVDNs, the extraction of the supplementary OSM data is skipped and γ_0 is set as 0.

The generated topology should be radial and connect all the households in the supply area, which is the operation requirement of LVDNs. To verify the number of cables under streets while generating a radial

Algorithm 1: Hierarchical Minimum Spanning Tree

Input: γ_0, D, D_s, $D_{s,1}$, $D_{s,2}$, D_{LV}, N_0, $\overline{I}$
Initial tree $\Gamma \leftarrow MST(D)$
if $\Gamma_w \notin D_{LV}$ **then**
  $D \leftarrow$ adjust the weights in D
  $\Gamma \leftarrow MST(D)$
end
if $\gamma_0 = 1$ **then**
  for $s \le S_0$ **do**
    Obtain I_s by Eq (1)
    if $I_s < \overline{I}$ **then**
      Sub-tree $\Gamma_0 \leftarrow MST(D_s)$
      Index I_0 of the non-root nodes in Γ_0
    else
      Sub-tree $\Gamma_0 \leftarrow MST(D_{s,1})$, $MST(D_{s,2})$
      Index I_0 of the non-root nodes in Γ_0
    end
  end
  $D_1 \leftarrow$ Update D based on I_0
end
Main tree $\Gamma \leftarrow MST(D_1)$
Graph topology $G \leftarrow \Gamma + \Gamma_0$
for $i \le N_t$ **do**
  Obtain I_s by Eq (1)
  if $I_s > \overline{I}$ **then**
    Remove l_{mi} in the cable
    Add an edge $l_{mi,3}$ between the nearest 3 degree point and the breakpoint
  end
end
Output: Graph topology: G^*

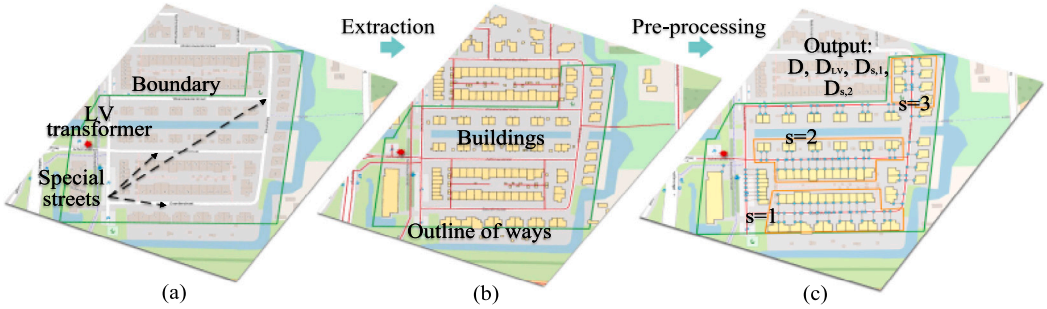


Fig. 2. Illustrative example of OSM data pre-processing: (a) is OSM with LV transformer and boundaries, (b) is raw OSM data in the LVDN and (c) is pre-processed OSM data.

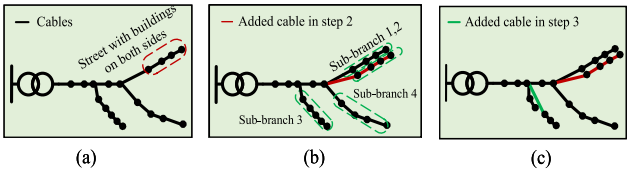


Fig. 3. Illustrative example of the output in each step: (a) is the initial graph topology, (b) is the revised topology in step 2 and (c) is the output of Algorithm 1.

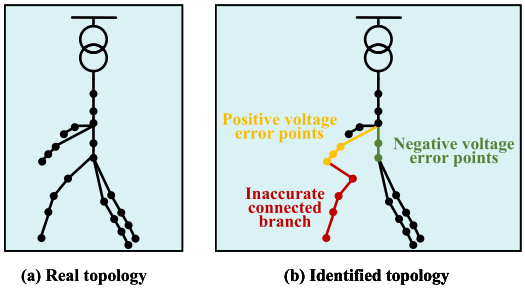


Fig. 4. Illustrative example for real and identification topology with inaccurate connection lines.

topology, a hierarchical minimum spanning tree (HMST) algorithm approach is constructed by incorporating a traditional MST algorithm and a peak demand-based refinement strategy, presented in Algorithm 1. The proposed HMST follows network planning principles and economic rationale: if street current does not surpass single line maximum capacity, deploying one cable is more cost-effective than two parallel cables. The input for the HMST algorithm consists of path matrices (D, D_s, $D_{s,1}$, $D_{s,2}$, D_{LV}), coordinates of N_0 and γ_0. S_0 is the number of streets with buildings on both sides. The traditional MST algorithm is adopted to generate a radial tree with the shortest length of cables, represented by $MST(\cdot)$ and the edge in the generated tree is represented by T_w. The peak demand-based refinement strategy is proposed to verify the number of cables under special streets based on the maximum capacity of underground cables and peak demand. The maximum load I_s of the street s is estimated using (1).

$$I_s = \frac{(r_g)^k \cdot N_s \cdot C_o \cdot P_{pe}}{3 \cdot cos\theta \cdot V_0}, \tag{1}$$

where r_g is the annual growth of demand and k is the planning period. C_o represents the concurrency for the N_s households located at street s, representing how many households reach peak load simultaneously, set as 0.46. P_{pe} is the average peak demand value. $cos\theta$ is the power factor of households.

Algorithm 1 can be divided into three main steps. In the first step, the weight in D is adjusted to ensure that all edges in the generated tree Γ are in D_{LV}. The output of this step is shown in Fig. 3(a). In the second step, for the streets with buildings on both sides, if the calculated maximum capacity I_s is larger than the rated maximum capacity $\bar{I}$ of the deployed cable. Two underground cables are assigned for this street (i.e., one cable for each side), and two sub-trees Γ_0 for the buildings located on each side of the street are obtained. If $I_s < \bar{I}$, one sub-tree is generated for all buildings located on this street. The topology Γ of the main feeders is generated based on the updated path matrix D_1. The initial graph topology of the LVDN is obtained by combining the sub-trees Γ_0 and the main tree Γ. The output of the second step is shown in Fig. 3(b). In the third step, the maximum load of all sub-branches in the topology is verified, and N_t is the number of the sub-branch. The degree of a node refers to the number of cables connecting the node. If the sub-branch is overloading, the connection line in the middle of the cable is removed, and the new connection line is placed between the breakpoint and the nearest 3-degree point (i.e., the intersections between the main street and the four sub-branches in Fig. 3(c)), which means that two cables are added for the front and back half of the street, respectively. Fig. 3(c) presents the generated graph topology G^* of the LVDN in the supply area, which includes the length of each cable and the connecting points.

On the other hand, the generated graph topology is close to the topology during the construction period, while the topology may have been updated due to component maintenance. Furthermore, the length of cables is only estimated using OSM data, whereas the length of deployed cables may be different from these values due to practical deployment factors. These problems will be illustrated and mitigated based on SM data in Sections 2.3 and 2.4.

2.3. Stage II: Topology reconstruction

The voltage drop ΔV^d_{mn} on line mn, total voltage drop ΔV^d_m from the transformer to bus m and voltage V_m at bus m are expressed as:

$$\Delta V^d_{mn,t} = \frac{P_{mn,t} + jQ_{mn,t}}{V_{m,t}}(r_{mn} + jx_{mn}) \tag{2}$$

$$\Delta V^d_{m,t} = \sum_{mn \in D^n_{LV}} \Delta V_{mn,t} \tag{3}$$

$$V_{m,t} = 1 - \Delta V^d_{m,t} \tag{4}$$

where P_{mn}, Q_{mn} are active and reactive power on line mn. r_{mn}, x_{mn} are resistance and reactance of line mn, respectively. D^n_{mn} is the line set from transformer to bus n.

When the output power of DERs is insufficient to reverse the direction of line flow (i.e., P_{mn} and Q_{mn} are positive), the voltage magnitude at node m decreases as its load increases. Therefore, if a branch is connected incorrectly in the identified topology, as illustrated by the line in Fig. 4(b), the load flowing through the green line will be reduced, leading to a diminished voltage drop. This means that the estimated voltage magnitude of the green nodes will be larger than the actual voltage magnitude (i.e., negative voltage magnitude residual).

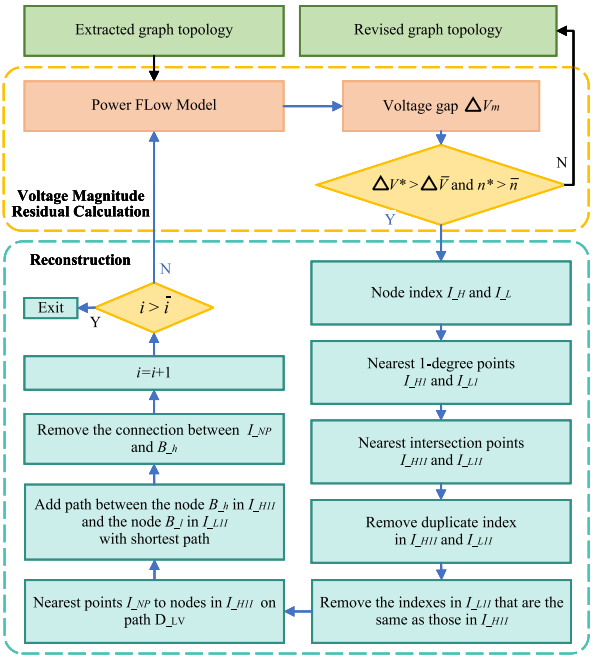


Fig. 5. Flowchart of proposed topology reconfiguration strategy. The upper part is voltage magnitude residual calculation based on a power flow model, the lower part is the street-to-street connection lines identification and reconstruction.

Conversely, for the yellow nodes, the estimated voltage magnitude will be smaller than the actual voltage magnitude and the voltage magnitude residual of these will be negative.

As previously discussed, the meshed streets in maps impact the final radial topology (i.e., the location of breakpoints). Thus, to obtain the accurate breakpoints of the underground cables, it is necessary to verify and revise the initial radial topology obtained from Stage I. A topology reconstruction strategy is introduced to revise the street-to-street connection lines with the assumption that two-time step SM data are available for each node, including voltage magnitude, net active power and net reactive power. The flowchart is illustrated in Fig. 5.

The ΔV^* represents the maximum voltage magnitude residual between the real topology and the extracted radial topology from Stage I. The voltage magnitude residual of each node is formulated as in (5). The threshold $\Delta\overline{V}$ of systemic error is pre-set by DSO or obtained from historical SM data. n^* represents the number of nodes with voltage residuals greater than $\Delta\overline{V}$ and its upper limit is $\bar{n}$. Meanwhile, to mitigate the impact of measurement errors, a margin η is introduced. The errors that are smaller than $\Delta\overline{V} - \eta$ are ignored.

$$\Delta V_m = \frac{1}{T}\sum_{t\in\mathcal{T}} (\mathrm{V}^0_{\mathrm{m,t}} - \mathrm{V}_{\mathrm{m,t}}), \quad m \in \mathcal{N} \tag{5}$$

where m represents the index of the nodes $\mathcal{N}$, T represents the dimension of time-series voltage data, t is the index of time, and V^c is the calculated voltage amplitude by the power flow model based on the extracted graph topology obtained from Stage I.

As shown in Fig. 5, the process of topology reconfiguration consists of voltage residual calculation and connection line modification, which is summarized as follows:

(1) Calculate the voltage residual ΔV^* and n^* based on the extracted graph topology and limited SM data.
(2) Check the topology based on $\Delta\overline{V}$, ΔV^*, n^* and $\bar{n}$. If ΔV^* is larger than $\Delta\overline{V}$ and n^* is larger than $\bar{n}$, the index I_H and index I_L of nodes with positive nodes (e.g., the yellow points in Fig. 4(b)) and negative residuals (e.g., the green points in Fig. 4(b)) are recognized, respectively. The nodes with positive errors indicate that they are connected to one cable with an additional load, resulting in a lower voltage magnitude than the actual value. If not, the extracted topology is the final topology. These two steps are illustrated in the orange dashed box in Fig. 5.
(3) Recognize the indexes I_{H1} and I_{L1} of the terminal node (1-degree node) of the cables with voltage residual and recognize the indexes I_{H1I} and I_{L1I} of the intersection nodes nearest to these 1-degree nodes.
(4) Identify the indexes in I_{L1I} that are the same as the indexes in I_{H1I}, and remove them from I_{L1I}. Steps 3 and 4 are shown as the right part of the bright blue dashed box in Fig. 5.
(5) Revise connection lines. A new connection line is constructed between the node pairs in I_{H1I} and I_{L1I} with the shortest paths, i.e., one node B_h in I_{H1I} and one node B_l in I_{L1I}. The connection line between node B_h and its upstream node is removed, avoiding the meshed structure.
(6) If the number of iterations reaches the threshold $\bar{i}$, exit the reconstruction and label the extracted graph topology as final. Otherwise, return to Step 1. The left part of the bright blue dashed box in Fig. 5 depicts the steps 5 and 6.

After topology reconstruction, the location of the breakpoints in the topology is defined, and the modified topology depicts the fundamental connection information of LVDNs.

2.4. Stage III: Topology optimization

Although the network topology obtained from Stage I and II is a feasible one, the length of deployed LV cables may be different from the actual length due to practical deployment factors, such as the crossing-street lines at crossroads. Moreover, the network topology may have been updated due to component maintenance. To address this issue, most traditional topology identification approaches are developed with the assumption that the time series measurements of each household are available [27,28], such as the regression-based topology identification. However, there may be a large number of unmetered households, or the deployed SMs fail to provide data for a short period. In this situation, only incomplete and sparse SM datasets are available, and the traditional power flow formulation based on the obtained topology is infeasible. Thus, to optimize the length of cables based on incomplete SM data, three data-driven optimization models are constructed based on the SM datasets with different incomplete rates as illustrated in Fig. 6. In Fig. 6, the blue blocks represent the available SM data, and the red blocks represent the unavailable or missed SM data. The complete SM data contains the voltage magnitude and demand profiles of each household at each time step, as shown in Fig. 6(a), while the incomplete SM data only consists of the partial profiles or the daily maximum and minimum measurements, as depicted in Fig. 6(b) and (c), respectively. Compared to the incomplete SM dataset in Fig. 6(b), the incomplete SM dataset in Fig. 6(c) is not only sparser but also asynchronous, leading to extra challenges for the proposed topology optimization procedure, which is based on a mathematical programming formulation, as presented next.

The topology optimization problem aims to define the length of cables, and it is formulated as a power-flow-based mathematical formulation based on the model in [25]. The LVDN is assumed to be balanced and is modeled as a single-phase network. Based on the three SM datasets with different incomplete rates in Fig. 6, three data-driven optimization models are stated as:

- Case I: a mathematical formulation that considers complete time-series SM data (i.e., Fig. 6(a)) is constructed. In this formulation, the net active and net reactive power of households are considered as input parameters.
- Case II: given the sparse SM dataset in Fig. 6(b), a mathematical formulation is developed to deal with the unavailable SM data, which takes net reactive and net active power as variables rather than parameters.

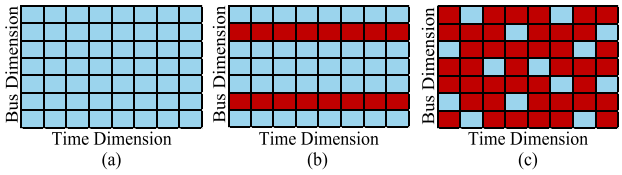


Fig. 6. Diagram of SM dataset with different incomplete rates: (a) complete dataset, (b) dataset with unmetered houses and (c) dataset with daily maximum and minimum values.

- Case III: when given the sparse and asynchronous SM dataset (i.e., the incomplete SM data in Fig. 6(c)), a mathematical formulation is stated by combining Case I and Case II. In this case, the daily SM data are utilized as the boundary for decision variables rather than variables.

2.4.1. Case I

To optimize the extracted topology from Section 2.3, we first assume that time series voltage magnitude and load profiles are accessible, and then the optimization problem is simplified into a single-level optimization model, given by formulation (6)–(14). The mathematical formulation for Case I aims to optimize the length of underground cables by minimizing the total voltage residuals of N nodes for the time horizon T. The voltage residual is defined as the square of the difference between the measured voltage magnitude $\mathrm{V}^{0}_{\mathrm{m,t}}$ and the estimated voltage magnitude $V_{m,t}$. The objective function is depicted in (6).

$$\min_{\hat{\mathcal{L}}} \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{N}} (V_{m,t} - \mathrm{V}^{0}_{\mathrm{m,t}})^2 \tag{6}$$

subject to:

$$\sum_{km\in\mathcal{L}} P_{km,t} + P^{s}_{m,t} - \sum_{mn\in\mathcal{T}} P_{mn,t} - \sum_{mn\in\mathcal{T}} \frac{1}{V^2_{m,t}} (P^2_{mn,t} + Q^2_{mn,t}) l_{mn} \widetilde{\mathrm{r}}_{\mathrm{mn}} = \mathrm{P}^{\mathrm{D0}}_{\mathrm{m,t}} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{7}$$

$$\sum_{km\in\mathcal{T}} Q_{km,t} + Q^{s}_{m,t} - \sum_{mn\in\mathcal{T}} Q_{mn,t} - \sum_{mn\in\mathcal{T}} \frac{1}{V^2_{m,t}} (P^2_{mn,t} + Q^2_{mn,t}) l_{mn} \widetilde{\mathrm{x}}_{\mathrm{mn}} = \mathrm{Q}^{\mathrm{D0}}_{\mathrm{m,t}} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{8}$$

$$V^2_{m,t} - V^2_{n,t} = 2(l_{mn}\widetilde{\mathrm{r}}_{\mathrm{mn}} P_{mn,t} + l_{mn}\widetilde{\mathrm{x}}_{\mathrm{mn}} Q_{mn,t}) - \frac{1}{V^2_{m,t}} (P^2_{mn,t} + Q^2_{mn,t})((l_{mn}\widetilde{\mathrm{r}}_{\mathrm{mn}})^2 + (l_{mn}\widetilde{\mathrm{x}}_{\mathrm{mn}})^2) \qquad \forall m,n \in \mathcal{N}, \forall mn \in \mathcal{L}, \forall t \in \mathcal{T} \tag{9}$$

$$\underline{V} \le V_{m,t} \le \overline{V} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{10}$$

$$\underline{\alpha}_{mn} \le l_{mn} \le \overline{\alpha}_{mn} \qquad \forall mn \in \mathcal{L} \tag{11}$$

where m and n represent the index of node $\mathcal{N}$, mn and km are the index of lines $\mathcal{L}$.

In Case I, the constant parameters (i.e., input data) are the SM data, including the historical voltage magnitude $\mathrm{V}_{\mathrm{m,t^0}}$, net active power $\mathrm{P}_{\mathrm{m,t^{D0}}}$ and net reactive power $\mathrm{Q}_{\mathrm{m,t^{D0}}}$ at node m at time t. The decision variables are the length rate of cable l_{mn}, and the variables in the power flow model (i.e., (7)–(10)), including line power flow $P_{mn,t}$, nodal voltage $V_{m,t}$, active power $P^{s}_{m,t}$ and reactive power $Q^{s}_{m,t}$ injected into the DNs. We assume that the LV transformer can provide sufficient active power and reactive power at the root node. Therefore, except for the root node, the active power $P^{s}_{m,t}$ and reactive power $Q^{s}_{m,t}$ at other nodes are set as 0. The power balance is ensured by constraints (7) and (8). Expression (9) models the voltage magnitude drop in the lines. The parameters $\widetilde{\mathrm{r}}_{\mathrm{mn}}$ and $\widetilde{\mathrm{x}}_{\mathrm{mn}}$ represent the impedance of the cable whose length is extracted from OSM data. The fourth item in constraints (7) and (8) represents the active power loss and reactive power loss. The variables $P_{km,t}$ and $P_{mn,t}$ represent the power flowing into and out of node m at time t, respectively. The voltage magnitude is limited by constraint (10).

The decision variables l_{mn} represent the ratio of optimized cable length and extracted length $\widetilde{l}_{mn}$ (i.e., the estimated cable length from OSM data), which is limited by constraint (11). The optimal solution is the ratio between the actual cable length and the extracted length. The parameters $\underline{\alpha}_{mn}$ and $\overline{\alpha}_{mn}$ are pre-set based on the quality of OSM data. Specifically, more accurate OSM data will provide more accurate information to estimate the length of underground cables, which means that the $|\underline{\alpha}_{mn} - \overline{\alpha}_{mn}|$ can be set to a smaller value. Meanwhile, a smaller value of $|\underline{\alpha}_{mn} - \overline{\alpha}_{mn}|$ means a smaller solution space, reducing the solving time of the before-presented mathematical formulation.

2.4.2. Case II

When there are several unmetered households, the formulation for Case II can be stated based on a state estimation model [29,30]. Compared to traditional methods in [31], the main advantage of the proposed mathematical formulation is that it does not need to check the location of the unmetered nodes and does not require that the parent or grandparent nodes of unmetered nodes are known. Compared to Case I, four additional variables are added to represent the unknown data of the unmetered households, including net active power $P^{D}_{m,t}$, net reactive power $Q^{D}_{m,t}$, the residuals of active power E^{P}_{m} and the residual of reactive power E^{Q}_{m}. The active power residuals and reactive power residuals are calculated by using expressions (14) and (15). Meanwhile, E^{P}_{m} and E^{Q}_{m} are added to the objective function, as shown in expression (12). The parameters w_v, w_p, and w_q are the weights for the three residuals, respectively. Besides, constraints (7) and (8) are revised as constraints (16) and (17). The complete mathematical formulation for Case II is then:

$$\min_{l_{mn}} \sum_{m\in N, t\in T} (w_v E^{V}_{m} + w_p E^{P}_{m} + w_q E^{Q}_{m}) \tag{12}$$

subject to: (9)–(11) and (13)–(17).

$$E^{V}_{m} = c_m (V_{m,t} - \mathrm{V}^{0}_{\mathrm{m,t}})^2 \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{13}$$

$$E^{P}_{m} = c_m (P^{D}_{m,t} - \mathrm{P}^{\mathrm{D0}}_{\mathrm{m,t}})^2 \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{14}$$

$$E^{Q}_{m} = c_m (Q^{D}_{m,t} - \mathrm{Q}^{\mathrm{D0}}_{\mathrm{m,t}})^2 \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{15}$$

$$\sum_{km\in\mathcal{L}} P_{km,t} + P^{s}_{m,t} - \sum_{mn\in\mathcal{T}} P_{mn,t} - \sum_{mn\in\mathcal{T}} \frac{1}{V^2_{m,t}} (P^2_{mn,t} + Q^2_{mn,t}) l_{mn} \widetilde{\mathrm{r}}_{\mathrm{mn}} = P^{D}_{m,t} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{16}$$

$$\sum_{km\in\mathcal{T}} Q_{km,t} + Q^{s}_{m,t} - \sum_{mn\in\mathcal{T}} Q_{mn,t} - \sum_{mn\in\mathcal{T}} \frac{1}{V^2_{m,t}} (P^2_{mn,t} + Q^2_{mn,t}) l_{mn} \widetilde{\mathrm{x}}_{\mathrm{mn}} = Q^{D}_{m,t} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{17}$$

where C is an index vector of the metered nodes, i.e., $[c_1, \ldots, c_m, \ldots, c_N]$. When node m is equipped with a smart meter, c_m is set as 1. Otherwise, c_m is set as 0, and the variables E^{V}_{m}, E^{P}_{m} and E^{P}_{m} all equal 0, which means that the residual of this household contributes nothing

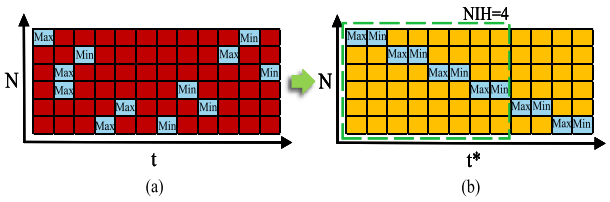


Fig. 7. Illustrative example of incomplete SM data reorganization: the sparse asynchronous dataset in (a) is converted into a stepped diagonal matrix in (b) with NIH set to 4.

to the objective function. Besides, it is flexible to pre-set parameters c_m according to the situation in LVDNs.

On the other hand, the error in the service line length is normally less than the error in the main feeder length. Thus, extra constraints for service lines may tighten the solution space for service lines. The constraint (18) is added to Case II. The limits for α and β are set according to the quality of OSM data, while the range of α should be set as larger than that of β.

$$\underline{\beta}_{mn} \le l_{mn} \le \bar{\beta}_{mn} \qquad \forall mn \in \mathcal{L}_s \tag{18}$$

where mn is the index of service line $\mathcal{L}_s$.

2.4.3. Case III

If only the daily maximum and minimum SM data (i.e., voltage magnitude and load profiles) are available, the input data for the mathematical formulation is sparse. As the maximum and minimum voltage magnitude do not occur simultaneously, the sparse SM data is also asynchronous, as shown in Fig. 6(c). To manage the impact of asynchronous data on the mathematical formulation, the SM data is reorganized into a stepped diagonal matrix, including the voltage magnitude and load profiles, as shown in Fig. 7. The pseudo-time horizon T^* is introduced as the x-axis for the reorganized SM data, which is defined as in (19).

$$T^* = \frac{2N \cdot T}{T_d}, \tag{19}$$

where T_d represents the dimension of the daily sample in Case-I.

The data point of each house is taken as the only known data in each sample (i.e., the blue block at each moment t^*) since the time of the daily maximum and minimum value is unknown. The remaining dimensions at each time t^* (i.e., the orange block) are then regarded as variables in the proposed mathematical formulation, constrained by the maximum and minimum values on the same day.

Constraint (10) in Case I is reformulated as constraint (21) in Case III. The active power and reactive power of each house are limited by constraints (22) and (23). The objective function of Case III is the same as that of Case I, but an additional weight (i.e., w_v) is introduced, as shown in the expression (20). The parameter c_m here represents whether the reorganized SM data of the house m is incorporated into the objective function. If the reorganized SM data of the houses 1–4 are integrated into the objective function (i.e., the blocks located in the green dashed box in Fig. 7(b)) then $c_{1,2,3,4} = 1$. Thus, as more nodes are integrated into the objective function, the reorganized SM data becomes more sparse. Additionally, a longer time horizon results in wider steps, consequently leading to sparser data. The number of integrated houses (NIH) is the sum of vector C, as in expression (24). Thus, a smaller time horizon T is adopted to decrease the sparseness of incomplete data.

The complete mathematical formulation for Case III is as follows:

$$\min_{\hat{\mathcal{L}}} \sum_{t\in\mathcal{T}} w_v \sum_{m\in\mathcal{N}} c_m (V_{m,t} - \mathrm{V}^0_{\mathrm{m,t}})^2 \tag{20}$$

subject to: (9), (11), (16)–(17) and (21)–(23).

$$\underline{V}^0_{m,t} \le V_{m,t} \le \overline{V}^0_{m,t} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T}^* \tag{21}$$

$$\underline{P}^{D0}_{m,t} \le P^D_{m,t} \le \overline{P}^{D0}_{m,t} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T}^* \tag{22}$$

$$\underline{Q}^{D0}_{m,t} \le Q^D_{m,t} \le \overline{Q}^{D0}_{m,t} \qquad \forall m \in \mathcal{N}, \forall t \in \mathcal{T}^* \tag{23}$$

$$NIH = \sum_{m\in N} c_m \tag{24}$$

where $\underline{V}^0_{m,t}$, $\overline{V}^0_{m,t}$, $\underline{P}^{D0}_{m,t}$, $\overline{P}^{D0}_{m,t}$, $\underline{Q}^{D0}_{m,t}$ and $\overline{P}^{D0}_{m,t}$ are the daily maximum and minimum SM data.

3. Case of study

In this section, the feasibility of the proposed topology generation approach is verified on three actual LVDNs in the Netherlands. The real topologies are illustrated in Fig. 8(a), (b), and (c), respectively, which are obtained from [32]. The base three-phase voltage is 0.4 kV. The extraction and pre-processing of raw OSM data is conducted in QGIS 3.28.1. The proposed hierarchical minimum spanning tree algorithm and all mathematical formulations are implemented in Python and Pyomo. All mathematical formulations are solved using the IPOPT solver. The time-series load profiles for each household are selected and scaled from reference [33], and the $cos\theta$ is set at 0.95 for each household. The voltage magnitude profiles are generated by using a PF model [25] and the real topologies. The voltage amplitude is limited to [0.90, 1.05] p.u., and the cable length is restricted to be within the range $[0.2 \cdot \tilde{l}_{mn}, 4 \cdot \tilde{l}_{mn}]$ (i.e., $\underline{\alpha}_{mn}$ and $\bar{\alpha}_{mn}$ are set at 0.2 and 4). For service lines, $\underline{\beta}_{mn}$ and $\bar{\beta}_{mn}$ are set as 0.7 and 1.3, respectively. Based on the preliminary results, the voltage residuals contribute more to the optimization problem, so w_v should be set much larger than the other two weights. The weights w_v, w_p and w_q in the objective function in (12) are set as 100, 1, and 1, respectively.

3.1. Graph topology generation

Different assumptions and data requirements make it challenging to directly compare the proposed approach with existing methods. The primary objective of the proposed approach is to generate a feasible near-real topology for low-voltage distribution networks. Therefore, this sub-section focuses on comparing the topology extracted by the proposed method with the actual topology for a number of case studies in The Netherlands.

Two residential LVDNs and one urban LVDN are selected to test the proposed HMST approach. The residential LVDN in Fig. 8(a) consists of 54 households and 52 nodes on the main feeders, which is named LV-52. The residential LVDN in Fig. 8(b) consists of 59 households and 62 nodes on the main feeders, which is named LV-62. The LVDN in Fig. 8(c) comprises 93 households and 95 nodes on the main feeders, which is named LV-95. The parameter γ_0 is set as 0 for LV-62 and LV-95, while it is set as 1 for LV-52. There are two streets with households on both sides in LV-52. The OSM data pre-processing for LV-52 is depicted in Fig. 9. Compared to the other two LVDNs, six sub-matrices (i.e., D_1, $D_{1,1}$, $D_{1,2}$, D_2, $D_{2,1}$, and $D_{2,2}$) are extracted from matrix D_0. Moreover, there is a circular road in LV-95 and LV-52, so the connection lines among households are not only subjected to the outline of streets but also limited by the physical constraint (i.e., the path matrix D_{LV}).

The generated graph topologies of the three LVDNs are shown in Fig. 8(d), (e), and (f). The generated topologies of LV-62 and LV-95 are very close to the actual topologies. However, compared to Fig. 8(a), there is an inaccurate connection line at the 39th node in the generated topology of LV-52 (i.e., Fig. 8(d)), which is caused by GIS data. Specifically, the path between the 39th node and the 31st node is smaller 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. This inaccurate connection line l_{39-31} should be revised by the suggested topology reconstruction strategy in the second stage.

Table 3 presents the total length of the main feeders and the number of nodes with different degrees. There is a difference between the estimated and real lengths of cables in the LV-52 and LV-62 networks, which is induced by missing OSM data (e.g., the missed houses and

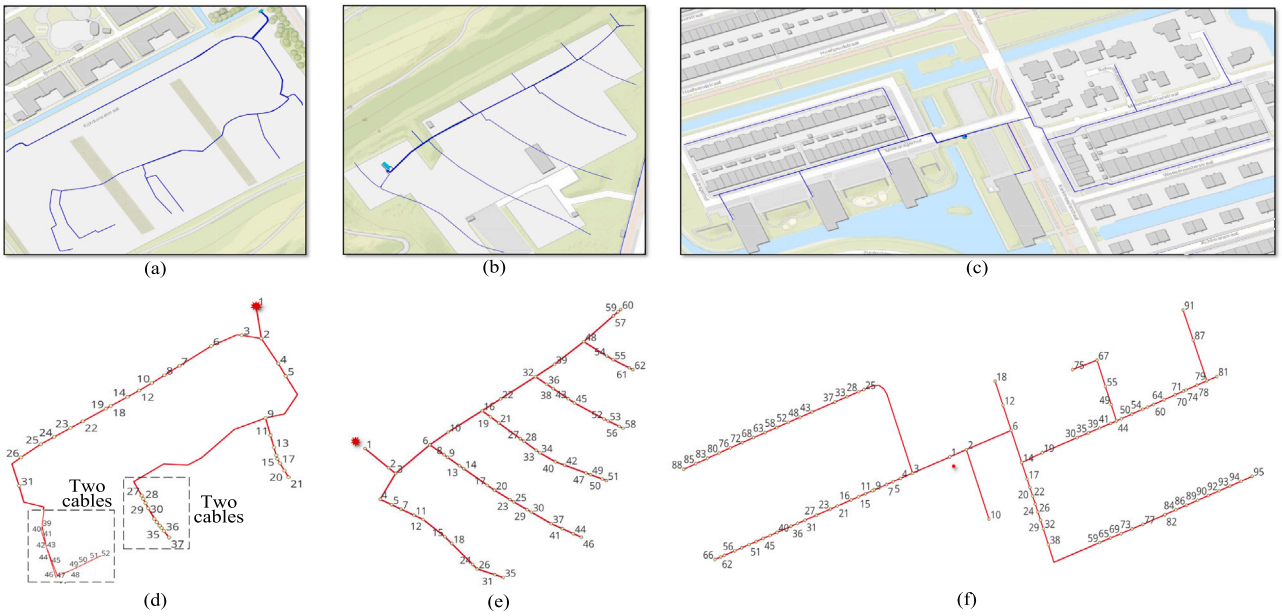


Fig. 8. Topology for (a) actual LV-52, (b) actual LV-62, (c) actual LV-95, (d) extracted LV-52, (e) extracted LV-62 and (f) extracted LV-95.

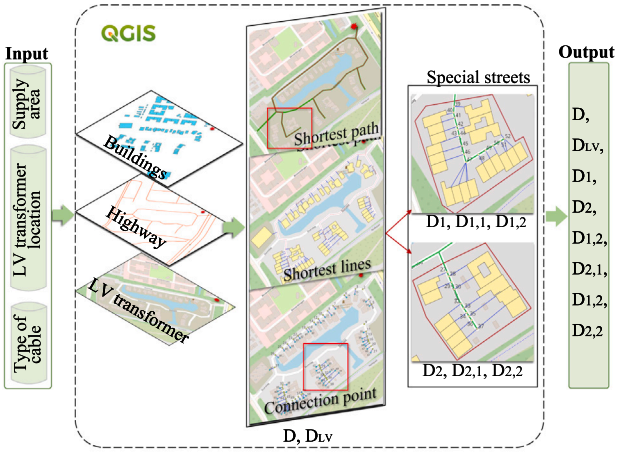


Fig. 9. Pre-processing of raw OSM data in LV-52: the yellow blocks represent buildings, the green lines are the outline of streets, and the straight blue lines represent the service lines.

the missed data of streets). The generated urban topology of the LV-95 network has higher accuracy due to the complete and precise OSM data. The shortest lines between buildings and main feeders are taken as service lines, which are directly extracted from OSM data. The extracted service lines are assumed to be connected accurately. The error in the main feeder length and service line length induced by inaccurate OSM data and deployment-related factors will be mitigated by making use of the proposed mathematical formulations for Cases I, II, and III.

3.2. Topology reconstruction

Based on preliminary results, the threshold $\Delta\overline{V}$ for these three systems is set as 0.005 p.u. and $\overline{n}$ is set as the number of the houses located at one of the sub-branches in the network. For the LV-62 and LV-95, the voltage magnitude residuals are 0.0029 p.u. and 0.0053 p.u. and n^* is 0 and 2, respectively, meaning that the connection information in these two systems is correct. However, the voltage magnitude residual

Table 3
Parameters of real topology and generated topology.

Network	Topology	Total length/m	Node degree		
			1	2	3
LV-52	Actual	796.91	7	41	5
	Generated	710.09	5	43	4
LV-62	Actual	730.81	12	45	11
	Generated	660.31	6	51	5
LV-95	Actual	1035.53	8	79	6
	Generated	1021.36	8	79	6

in LV-52 is 0.0108 p.u. and n^* is 20, indicating a connection error, which can be seen from the extracted graph topology in Fig. 8(d). The 39th household should be connected to the 27th household but is wrongly connected to the 31st household, which leads to a larger error. The LV-52 network and three modified LVDNs are used to test the proposed topology reconstruction strategy in this section and illustrated in Table 4. The parameter $\bar{n}$ in LV-52 is 7. The modified topology is used to analyze the impact of the number of sub-branches around the breakpoint on the topology reconstruction, i.e., relocating the breakpoints. The network LV-52 represents the network with one sub-branch located far away from the breakpoint, while the network LV-52-I and the network LV-52-II represent the network with one sub-branch located near the breakpoint. The network LV-52-III represents the network with two wrongly connected sub-branches around the breakpoint. Besides, the length of line l_{40-31} in the network LV-52-III is added 40 meters to make it close to the length of line l_{39-31} since the 40th household is next to the 39th household.

After applying the proposed topology reconstruction strategy, the revised topology and the voltage magnitude residual in each iteration are shown in Table 5. Given the pre-set voltage magnitude threshold, the extracted LVDN topology with only one wrongly connected sub-branch is revised to the correct one after one iteration. The voltage magnitude residual and parameter n^* in LV-52-III decreases with iterations, indicating that the extracted LVDN topology with two wrongly connected sub-branches is revised to the accurate topology after two iterations. Table 6 shows that the proposed topology reconstruction strategy efficiently identifies the location the of breakpoint and revises the wrongly connected sub-branches using data with errors. Compared

Table 4
Modified topology and extracted topology.

Network	True topology	Modified topology	Modified length/m	Extracted topology
LV-52	l_{9-27} l_{40-39} l_{39-27}		0 0 0	l_{9-27} l_{40-39} l_{39-31}
LV-52-I	l_{9-27} l_{40-39} l_{39-27}	l_{9-27}	−100 0 0	l_{9-27} l_{40-39} l_{39-31}
LV-52-II	l_{9-27} l_{40-39} l_{39-27}	l_{40-31}	0 +40 0	l_{9-27} l_{40-31} l_{39-31}
LV-52-III	l_{9-27} l_{40-39} l_{39-27}	l_{40-27}	0 +100 0	l_{9-27} l_{40-31} l_{39-31}

Table 5
Topology reconfiguration results.

Network	Iteration	Voltage residual/p.u.	n^*	Extracted topology	Revised topology
LV-52	0	0.0108	20	l_{39-31}	
	1	0.0001	0		l_{39-27}
LV-52-I	0	0.0108	9	l_{39-31}	
	1	0.0001	0		l_{39-27}
LV-52-II	0	0.0054	13	l_{40-31} l_{39-31}	
	1	0.0001	0		l_{40-31} l_{39-27}
LV-52-III	0	0.0108	20	l_{40-31} l_{39-31}	
	1	0.0089	15		l_{40-31} l_{39-27}
	2	0.0024	0		l_{40-27} l_{39-27}

Table 6
Topology reconfiguration for LV-52 under measurement error.

Error percentage	Iteration	Voltage residual/p.u.	n^*	Feasibility flag
0.2%	0	0.0112	19~23	1
	1	0.0016	0	
0.5%	0	0.0119	22~29	1
	1	0.0031	0	
1%	0	0.0142	26~31	1
	1	0.0063	1~4	
2%	0	0.0184	34~38	0

to the traditional topology reconfiguration approaches, the proposed strategy focuses on modifying the street-to-street connection lines in the graph topology instead of optimizing the grid structure based on SM data.

The impact of measurement errors in voltage magnitude impacts the feasibility of the proposed topology reconstruction strategy. According to the IEC 62053-21 standard [34], we considered four classes of SM, including 0.2%, 0.5%, 1% and 2%, which represent the maximum relative error compared to real data. The margin η is set as 0.2%. The experiments were executed five times, and the averaged voltage magnitude residual and the range of parameter n^* are summarized in Table 4. As the magnitude of error increases, both the averaged voltage magnitude residual and n^* show an upward trend, indicating that measurement errors affect the reconstruction process. For highly precise SMs (0.2%, 0.5%, and 1%), the proposed strategy remains feasible. However, it becomes infeasible when the error magnitude reaches 2% due to two main reasons: (1) Larger errors make it unable to identify the real potential connection points for inaccurate connection lines, resulting in IL and IH containing too many duplicated nodes. (2)

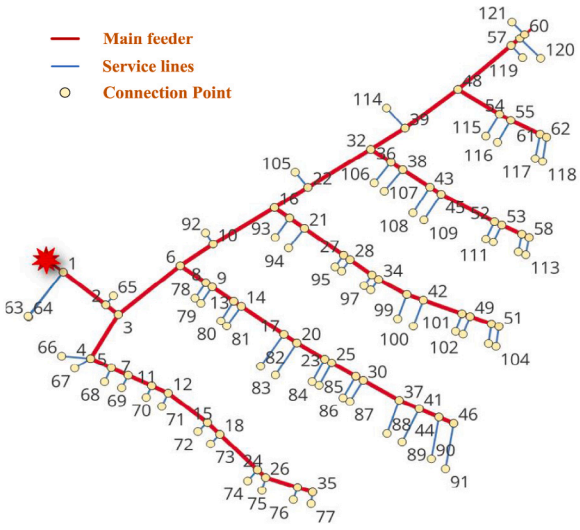


Fig. 10. Extracted graph topology with service lines of LV-62.

Larger errors may lead to the identification of incorrect potential connection points, triggering the method termination upon reaching the maximum iteration limit. Nevertheless, the proposed strategy ensures the accuracy of topology reconstruction when applied to datasets with smaller error magnitudes.

However, the length of the underground cables in the revised topology may be incorrect due to the cable replacement and inaccurate OSM data. In the next section, these errors will be mitigated in the topology optimization stage. Meanwhile, the service lines between houses and main feeders are integrated into the topology, while the above two stages only focus on the main feeders. The LV-62 with service lines is shown in Fig. 10.

3.3. Topology optimization

3.3.1. Case I

To analyze the impact of the amount of SM data, SM datasets with different time horizons (i.e., T) are taken as the input for the Case I mathematical formulation. The mean and maximum relative errors of length rate l_{mn} in three LVDNs are illustrated in Fig. 11. The mean relative error of estimated cable length is less than 0.2%, indicating that the proposed mathematical formulation effectively obtains the real length of cables based on complete SM data. As expected, the maximum relative errors for the LV-52, LV-62, and LV-95 networks are larger than 10%, 30% and 7%, respectively, if only SM datasets with shorter horizons are available. The maximum relative error of estimated cable length decreases by less than 6% when more SM data is accessible, such as one-day SM data with a 15-minute resolution. Meanwhile, high-dimension SM data increases the complexity of the Case-I, making it time-consuming to solve the proposed optimization model, which is depicted in Fig. 11(d). Nevertheless, when given a one-day SM dataset with a 15-minute resolution, the solving time is less than 1 min.

3.3.2. Case-II

An incomplete SM dataset with missing data from five unmetered households is taken as an example for Case II. The number of unmetered households with missing SM data in the LV-52, LV-62, and LV-95 networks are [72, 76, 80, 89, 102], [75, 78, 98, 101, 113], and [141, 153, 174, 176, 187], respectively. Fig. 12 illustrates the error in length ratio l_{mn} of all lines when using the proposed mathematical formulation for Case II, solved without and with constraint (18). Fig. 12 shows that including constraint (18) in the proposed mathematical formulation improves the general accuracy. Specifically, the maximum

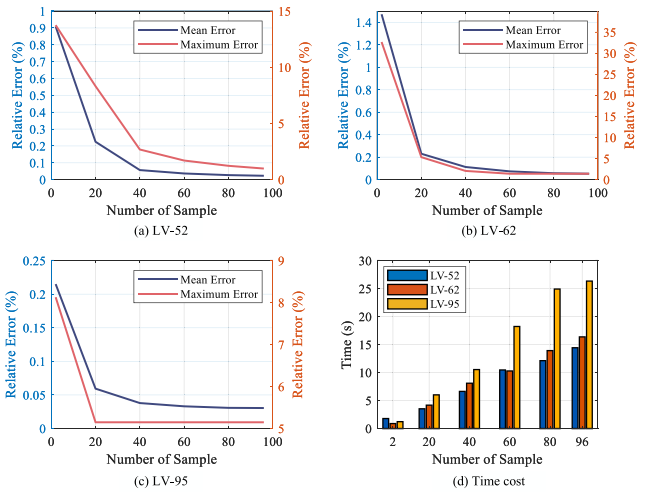


Fig. 11. Relative error of l_{mn} and calculation time in Case I. As shown in (a), (b) and (c), cable lengths are correctly estimated from complete data and the calculation time is less than 30 s (d).

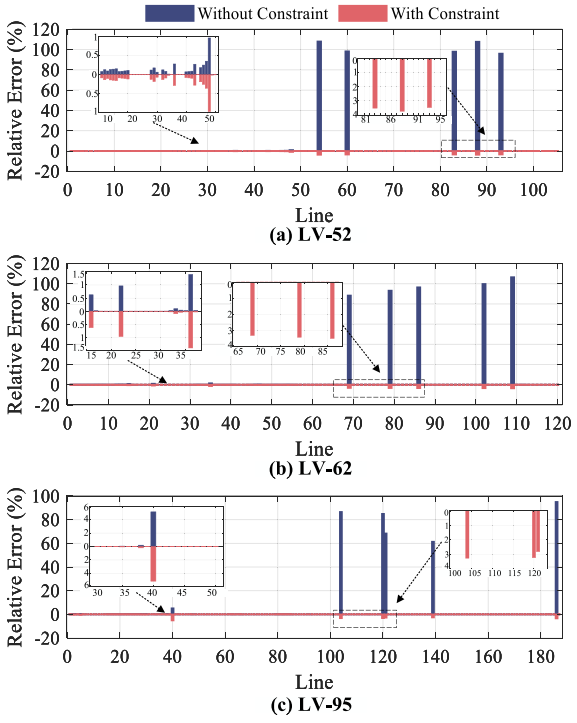


Fig. 12. Relative error of l_{mn} of each cable in Case-II. The red bars illustrate that constraint (18) for the service lines enhanced the accuracy.

relative error in the service lines in the LV-52 and LV-62 networks is below 4%, and that in the LV-95 network is below 6%. Thus, as expected, specific constraints for each type of cable tighten the solution space, increasing the accuracy of the estimation of underground cable length.

To analyze the impact of the unmetered houses on the performance of Case II, the time horizon $\mathcal{T}$ is set as 50 and 90, which means that there are 50 or 90 data samples (i.e., V, P, and Q) for each available household. Meanwhile, different unmetered rates R_{unm} in the LVDNs are set for the three LVDNs, defined as:

$$R_{unm} = \frac{N_{un}}{N_0} \times 100\% \tag{25}$$

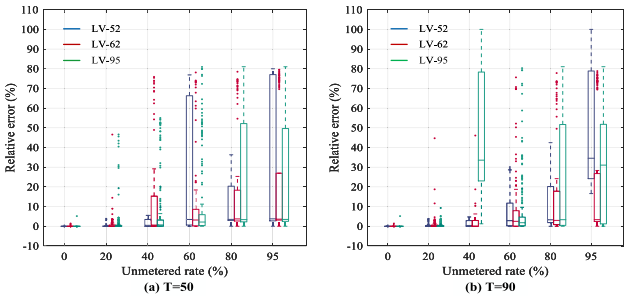


Fig. 13. Distribution of relative error of l_{mn} in Case-II under multiple unmetered rates and different time horizons The relative error experiences a rapid increase when R_{unm} exceeds 30%.

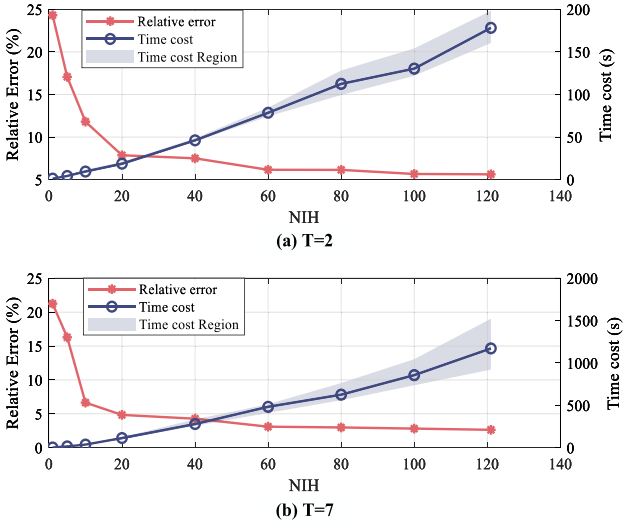


Fig. 14. Mean relative error of l_{mn} and calculation time in Case-III under multiple NIH and different time horizons. The intersection of the two curves falls within the NIH range of 20 to 60.

Here, R_{unm} represents the ratio of the N_{un} houses without available smart meters to the total N_0 houses in the LVDNs. The higher the ratio of unmetered nodes, the fewer available SM data, resulting in a sparse input SM dataset.

Fig. 13 shows the relative error of l_{mn} under different unmetered rates and time horizons. Only the relative errors that are lower than 100% are shown in this paper. The points represent the larger error, and the boxes depict the interval of the rest of the relative error. Compared to the LV-95 network, the cable length in the LV-52 and LV-62 networks is accurately identified when using less sparse input SM data. Specifically, when T is set as 90 and R_{unm} is up to 40%, the relative error in cable length for the LV-52 network remains below 10%, and below 20% for the LV-62 network. For the LV-95 network, the relative error in the cable length is less than 20% with T set as 90 and R_{unm} as 20%. These errors are located in the acceptable interval, according to [35]. However, the relative error increases rapidly as the unmetered rate increases, especially for networks LV-62 and LV-95. Given the same unmetered ratio R_{unm}, more SM data (i.e., a larger time horizon) improves the accuracy of the proposed mathematical formulation in Case II. In conclusion, the location of the unmetered houses also impacts the feasibility of the proposed mathematical formulation and its accuracy.

3.3.3. Case III

For Case III, the proposed mathematical formulation is tested on the LV-62 network. The time horizon $\mathcal{T}$ is set as 2 and 7. A larger $\mathcal{T}$ leads

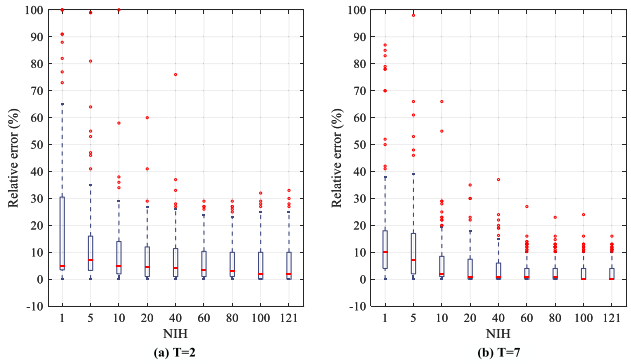


Fig. 15. Distribution of relative error of l_{mn} in Case-III under multiple HIN and different time horizon. The relative error experiences a rapid decrease when NIH exceeds 10%.

to a wider ladder in the reorganized data format, which also affects the accuracy of the proposed mathematical formulation. Meanwhile, to analyze the impact of NIH, this is set at [1, 5, 10, 20, 40, 60, 80, 100, 121]. The mean relative error of l_{mn} and the average calculation time are depicted in Fig. 14, which depicts that the solving time of the model increases with the increase of NIH. The complexity of the optimization model concerning NIH is approximately between O(n) and O (n^2). The intersection of the two lines represents the local optimal point, meaning that the actual lengths are identified with lower relative errors and less computational time is required. In Fig. 14, the local optimal NIH is located between 20 and 60, where the mean relative error of cable length remains below 10% and the computation time is below ten minutes. On the other hand, solving time is influenced not only by the complexity of the model but also by factors such as the location of nodes whose voltage residuals are integrated into the objective function, hardware limitations, long-term computational processes on a single laptop, and other factors. Thus, the intersection could be used by DSO to set a proper NIH value according to the accuracy requirements. The relative error distribution under the above scenarios is illustrated in Fig. 15. The red points represent the larger error, and the boxes depict the interval of the rest of the relative error. When NIH is set between 20 and 60, the relative errors predominantly distribute within 0%–10%, with a few falling within the 10% to 80% interval. Thus, the proposed mathematical formulation in Case III is feasible, given the well-designed parameters NIH and T.

Compared to Case I, the datasets used in Case II and III are more sparse and asynchronous, which leads to the error magnitude increases, as expected. Based on Figs. 11, 13 and 15, the longer the period of the available SM data and the more households with SM installed, the closer the estimated cable length is to the true value, i.e., the smaller length error. The ideal situation is that all households are assumed to have smart meters and can provide more than one month of high-resolution SM data, which is the assumption of most papers. Moreover, in case III, the longer period of the available SM data will lead to sparser transformed data (as shown in Fig. 7), which will increase the calculation burden, as shown in Fig. 14. Thus, the models in cases I to I can be selected and the parameters in the models can be set according to the period of the available data and the number of available SM, to ensure that the estimated cable length is close to the true value.

4. Conclusion

The network topology is significant for the efficient operation and planning of distribution networks, while it is challenging to obtain accurate topologies due to missing recordings, high-frequency maintenance, and user phase shifting. This paper proposes a topology identification approach for LVDNs with high-proportion underground cables based on graph topology generation, topology reconstruction, and topology optimization. The proposed approach is tested in three actual LVDNs in the Netherlands and multiple incomplete SM datasets. According to the obtained results, the proposed HMST algorithm can generate a graph topology with an accurate number of cables for each street. However, inaccurate street-to-street connections in the graph topology can be found, induced by mesh streets. These inaccuracies are then successfully revised by the topology reconstruction strategy. The generation of graph topology only relies on open map data and SM data, which makes it more flexible than existing approaches. Nevertheless, the inaccurate OSM data and the deployment environment lead to inaccurate cable length. Considering that only metered houses or daily maximum and minimum SM data are available, three data-driven optimization models were stated. The results showed that the proposed mathematical formulations successfully decreased the error in cable length. Moreover, the results also illustrated the minimal amount of SM data needed to minimize the error of cable length under multiple incomplete SM datasets.

CRediT authorship contribution statement

Dong Liu: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Juan S. Giraldo:** Writing – review & editing. **Peter Palensky:** Funding acquisition. **Pedro P. Vergara:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by China Scholarship Council (CSC) (Grant No. 202206130017).

Data availability

Data will be made available on request.

References

[1] Antić T, Capuder T, Bolfek M. A comprehensive analysis of the voltage unbalance factor in PV and EV rich non-synthetic low voltage distribution networks. Energies 2020;14(1):117.
[2] Cavraro G, Arghandeh R. Power distribution network topology detection with time-series signature verification method. IEEE Trans Power Syst 2017;33(4):3500–9.
[3] Zhang H, Zhao J, Wang X, Xuan Y. Low-voltage distribution grid topology identification with latent tree model. IEEE Trans Smart Grid 2022;13(3):2158–69.
[4] Costa LF, Giraldo JS, Castro CA. Identification and correction of transmission line parameter errors using SCADA and synchrophasor measurements. Int J Electr Power Energy Syst 2022;135:107509.
[5] Birchfield AB, Xu T, Gegner KM, Shetye KS, Overbye TJ. Grid structural characteristics as validation criteria for synthetic networks. IEEE Trans Power Syst 2017;32(4):3258–65.
[6] Koirala A, Suárez-Ramón L, Mohamed B, Arboleya P. Non-synthetic European low voltage test system. Int J Electr Power Energy Syst 2020;118:105712.
[7] He X, Qiu RC, Ai Q, Zhu T. A hybrid framework for topology identification of distribution grid with renewables integration. IEEE Trans Power Syst 2021;36(2):1493–503.
[8] Liao Y, Weng Y, Liu G, Rajagopal R. Urban MV and LV distribution grid topology estimation via group lasso. IEEE Trans Power Syst 2019;34(1):12–27.
[9] Shah P, Zhao X. Network identification using μ-PMU and smart meter measurements. IEEE Trans Industr Inform 2022;18(11):7572–86.
[10] Wang X, Zhao Y, Zhou Y. A data-driven topology and parameter joint estimation method in non-pmu distribution networks. IEEE Trans Power Syst 2024;39(1):1681–92.
[11] Liu Y, Wang J, Wang P. Hybrid data-driven method for distribution network topology and line parameters joint estimation under small data sets. Int J Electr Power Energy Syst 2023;145:108685.

[12] Yu J, Weng Y, Rajagopal R. PaToPaEM: A data-driven parameter and topology joint estimation framework for time-varying system in distribution grids. IEEE Trans Power Syst 2019;34(3):1682–92.
[13] Cunha VC, Freitas W, Trindade FC, Santoso S. Automated determination of topology and line parameters in low voltage systems using smart meters measurements. IEEE Trans Smart Grid 2020;11(6):5028–38.
[14] Zhao L, Liu Y, Zhao J, Zhang Y, Xu L, Xiang Y, Liu J. Robust PCA-deep belief network surrogate model for distribution system topology identification with DERs. Int J Electr Power Energy Syst 2021;125:106441.
[15] Duan Y, Wang C, Zhou W. Topology modeling of distribution network based on open-source GIS. In: 4th international conference on electric utility deregulation and restructuring and power technologies. DRPT, 2011, p. 527–30.
[16] Ali M, Macana CA, Prakash K, Islam R, Colak I, Pota H. Generating open-source datasets for power distribution network using OpenStreetMaps. In: 9th international conference on renewable energy research and application. ICRERA, 2020, p. 301–8.
[17] Nasirifard P, Rivera J, Zhou Q, Schreiber KB, Jacobsen H-A. A crowdsourcing approach for the inference of distribution grids. In: Proceedings of the ninth international conference on future energy systems. 2018, p. 187–99.
[18] Mateo Domingo C, Gomez San Roman T, Sanchez-Miralles A, Peco Gonzalez JP, Candela Martinez A. A reference network model for large-scale distribution planning with automatic street map generation. IEEE Trans Power Syst 2011;26(1):190–7.
[19] Çakmak HK, Janecke L, Weber M, Hagenmeyer V. An optimization-based approach for automated generation of residential low-voltage grid models using open data and open source software. In: Proc. IEEE power energy soc. innov. smart grid technol. conf.. ISGT, 2022, p. 1–6.
[20] Kays J, Seack A, Smirek T, Westkamp F, Rehtanz C. The generation of distribution grid models on the basis of public available data. IEEE Trans Power Syst 2017;32(3):2346–53.
[21] Mateo C, Prettico G, Gómez T, Cossent R, Gangale F, Frías P, Fulli G. European representative electricity distribution networks. Int J Electr Power Energy Syst 2018;99:273–80.
[22] Pisano G, Chowdhury N, Coppo M, Natale N, Pilo F. Synthetic models of distribution networks based on open data and georeferenced information. Energies 2019;12(23):4500.
[23] Sarajlić D, Rehtanz C. Low voltage benchmark distribution network models based on publicly available data. In: Proc. IEEE power energy soc. innov. smart grid technol. conf.. ISGT, 2019, p. 1–5.
[24] Grzanic M, Flammini MG, Prettico G. Distribution network model platform: A first case study. Energies 2019;12(21):4079.
[25] Vergara PP, López JC, Rider MJ, Da Silva LC. Optimal operation of unbalanced three-phase islanded droop-based microgrids. IEEE Trans Smart Grid 2017;10(1):928–40.
[26] Watson JD, Welch J, Watson NR. Use of smart-meter data to determine distribution system topology. J Eng 2016;2016(5):94–101.
[27] Korres GN, Manousakis NM. A state estimation algorithm for monitoring topology changes in distribution systems. In: In proc. IEEE power energy soc. gen. meeting. 2012, p. 1–8.
[28] Karimi HS, Natarajan B. Joint topology identification and state estimation in unobservable distribution grids. IEEE Trans Smart Grid 2021;12(6):5299–309.
[29] Jia K, Yang Z, Zheng L, Zhu Z, Bi T. Spearman correlation-based pilot protection for transmission line connected to PMSGs and DFIGs. IEEE Trans Ind Inf 2020;17(7):4532–44.
[30] Acurio BAA, Barragán DEC, López JC, Grijalva F, Rodríguez JC, da Silva LCP. State estimation for unbalanced three-phase AC microgrids based on mathematical programming. In: Proc. IEEE power energy soc. innov. smart grid technol. conf.. ISGT, 2023, p. 1–5.
[31] Deka D, Backhaus S, Chertkov M. Structure learning and statistical estimation in distribution networks-part II. 2015, arXiv:1502.07820.
[32] Liggingsdata kabels en leidingen. 2023, https://www.stedin.net/zakelijk/opendata/liggingsdata-kabels-en-leidingen. [Accessed August 2023].
[33] Schneider KP, Mather B, Pal BC, Ten C-W, Shirek GJ, Zhu H, Fuller JC, Pereira JLR, Ochoa LF, de Araujo LR, et al. Analytic considerations and design basis for the IEEE distribution test feeders. IEEE Trans Power Syst 2017;33(3):3181–8.
[34] García S, Mora-Merchán JM, Larios DF, Personal E, Parejo A, León C. Phase topology identification in low-voltage distribution networks: A Bayesian approach. Int J Electr Power Energy Syst 2023;144:108525.
[35] Chérot G, Latimier RLG, Sanchez F, Ahmed HB. Misestimation of impedance values within a distribution network optimal power flow. In: Proc. IEEE belgrade powerTech. 2023, p. 1–6.