SMART METERS PHASE IDENTIFICATION FOR TOPOLOGY VERIFICATION: PRACTICAL CHALLENGES AND INSIGHTS FROM A CASE STUDY

Thibaut Théate[1], Laurine Duchesne[1], Adrien Leerschool[1], Alireza Bahmanyar[1*], Simon Gerard[2], Thomas Wehenkel[2], Damien Ernst[2,3],

[1]Haulogy, Rue Bonry 53D, 4120 Neupré, Belgium
[2]RESA, Boulevard d'Avroy 38, 4000 Liège, Belgium
[3]Department of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, Belgium
**alireza.bahmanyar@haulogy.net*
{thibaut.theate, alireza.bahmanyar, laurine.duchesne}@haulogy.net
{simon.gerard, thomas.wehenkel}@resa.be
dernst@uliege.be

Keywords: Distribution System Operator, Smart meters, Power grid topology, Pearson correlation.

Abstract

For historical reasons, many distribution system operators (DSOs) do not have complete and accurate information about their networks. In particular, the topology of the distribution power grid, which is required to carry out numerous analyses necessary to the energy transition, is generally imprecise. This research work presents a novel approach to verify the distribution network topology on the basis of imperfect data. More precisely, the dynamic data retrieved from a partial coverage of smart meters are utilised for that purpose. Relying on the clustering of electrical phases on the basis of the correlation between voltage time series, the algorithmic solution proposed has the key advantage of yielding interpretable results, in contrast to black-box models. Promising results are reported for a real test case: an area managed by RESA, a Belgian DSO, with a coverage of smart meters inferior to 20%. Numerous ambiguous situations in the dataset of the DSO have been correctly identified by the algorithm, whose output was perfectly logical from a human's perspective. To finish with, the interpretability of the results helped to identify ideas to be investigated as future work in order to further improve the algorithmic solution presented.

1 Introduction

Distribution System Operators (DSOs) are expected to play an important role in the ongoing energy transition. Indeed, the electrification of many uses is impossible without a major update of the electrical distribution network. Nevertheless, effectively upgrading the power grid is far from being trivial. This key task requires up-to-date and in-depth knowledge of the distribution network, an information that many DSOs currently lack. In fact, the data available to the DSO are generally incomplete, inaccurate, and outdated. This situation motivates the ongoing research on topology discovery for the distribution power grid. Among others, having access to an accurate topology of the network enables valuable analyses, including the computation of hosting capacity and the prioritisation of investments.

In the context of DSOs, topology discovery involves the identification of the path of electricity for each customer connected to the distribution power grid. In other words, such an exercise boils down to associating each meter to the right feeder. The scientific literature includes interesting studies focused on discovering the topology of distribution power grids using imperfect data, see e.g. [1]-[3]. However, even the best algorithms can still be affected by inaccurate input data, leading to possible errors in the results.

In the energy transition process, DSOs need to leverage all available resources to overcome such challenges effectively. In this context, this research work suggests a verification of the reconstructed network topology by leveraging dynamic data collected by smart meters, whether they are fully or partially deployed. More precisely, the proposed method serves as a

processing block besides topology identification, providing updated information on the connection phases of the customers equipped with smart meters. In contrast to the previous studies on the subject [4]-[5], this research work focuses on a practical scenario where transformer measurements are not always available, and where the deployment of smart meters is limited. Moreover, designed on the basis of an intuitive methodology, the algorithmic solution proposed has the key advantage of yielding interpretable results, as opposed to black-box models.

2. Problem formalisation

As previously explained, the problem studied in this research work can be summarised as the verification of the distribution network topology on the basis of dynamic data measured by smart meters. In order to achieve this objective, the present scientific article suggests, for each feeder, clustering the electrical phase(s) measured by all the smart meters presumably assigned to that feeder.

More formally, the input information available includes two key elements:

- The topology reconstructed on the basis of another methodology (e.g. with static data), which includes the set of pairs {meter, feeder};
- The voltage time series data collected by all types of smart meters and feeders over a certain time period.

As output, the algorithmic solution has to provide a set of triplets {meter, phase, feeder}. This output will also highlight the potential errors identified in the previous topology.

A major complexity of the topology problem studied is the unavailability of a relevant ground truth, for obvious reasons. Therefore, no quantitative criteria can be defined to rigorously evaluate the performance of algorithmic solutions. Instead, performance assessment has to be carried out qualitatively from a human perspective.

3. Methodology

The core idea of the novel methodology proposed is, for each feeder of the distribution power grid, to cluster the voltage time series into four different clusters: three for the electrical phases and one for the potential outliers. The Pearson correlation coefficient has been selected as the distance criterion between two time series. This particular criterion has been empirically preferred over the Spearman and Kendall approaches, which have also been tested. If this correlation is below a certain threshold, it is an indication that the smart meter analysed may potentially not be associated to the feeder initially assigned.

Not systematically having the transformer measurements (three phases for each feeder), the main complexity with this methodology is identifying a relevant reference whose three electrical phases will serve as starting points for the three clusters. When there is a single three-phase smart meter assumed within the feeder, it can serve as the reference. When there are additional three-phase meters available, the cross-correlation matrix between these smart meters is computed. The reference selected is the three-phase smart meter whose cumulative correlation is the highest. But, in the case of a very poor coverage of smart meters or lack of three-phase ones, such a representative reference may not always be available. In that case, the algorithmic solution proposed is unfortunately not applicable for these feeders.

The correlation threshold is an important parameter of the proposed methodology. As illustrated in the results section, the first experiments conducted highlighted the necessity to consider a function for this threshold instead of a constant. Indeed, the correlation has been observed to be strongly dependent on three elements in practice:

- The length of the power line connecting the reference to the smart meter analysed;
- The number of production/consumption points in between the reference and the smart meter analysed;
- The structure of the lines connecting the reference to the smart meter (e.g., division, loop, end of line).

While the first two are taken into consideration in the algorithmic solution proposed, the latter is left as future work. More formally, the correlation threshold is defined as follows. Let C_T be the correlation threshold, it can be mathematically expressed as the function:

$$C_T = \max\left[C, C_{min}\right], \quad \text{where } C = C_{max} - \alpha \times L - \beta \times N,$$

with the following parameters:

- $C_{min} \in [0, 1]$ being the minimum value for the correlation threshold;
- $C_{max} \in [0, 1]$ being the maximum value for the correlation threshold;
- $L \geq 0$ being the length of the power line connecting the reference and the smart meter analysed;
- $N \in \mathbb{N}$ being the number of production/consumption points in between the reference and meter analysed;
- $\alpha \geq 0$ and $\beta \geq 0$ being parameters quantifying the expected decrease in correlation with both the length of the power line and the number of consumers in between the reference and the smart meter analysed.

If the correlation between a smart meter and the reference three-phase meter is inferior to the threshold, this particular smart meter is labelled as a potential outlier. Once the clustering operation has been performed for all feeders, the potential outliers undergo an additional analysis. For each smart meter, the correlation with the three clusters (phases) of the neighbouring feeders within a certain range is computed. If the maximum correlation is superior to the corresponding threshold, a suggestion is issued to associate the smart meter to this alternative feeder. Otherwise, the smart meter is definitively tagged as an outlier requiring special treatment.

The pseudocode in Fig. 1 outlines the main steps of the smart meters clustering operation, when measurements are available at the transformation points. Moreover, Fig. 2 presents the approach for generating suggestions of topology correction.

Algorithm 1 Intuitive clustering of smart meters

Loop through the K feeders composing the distribution power grid.
for $k = 1$ **to** K **do**
  Get the list of meters connected to feeder k: L_M.
  Set as references of three clusters the three phases measured at feeder k.
  while `length`$(L_M) \geq 0$ **do**
    Get a meter M from the list L_M.
    Build a $3 \times n$ correlation matrix between the n phases of M and 3 references.
    Assign each phase of M to a cluster, in decreasing order of correlation C_n.
    Compute the average correlation of meter M: $C_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} C_i$.
    Compute the correlation threshold $C_{\text{T}} = \max\ [C_{\max} - \alpha \cdot L - \beta \cdot N,\ C_{\min}]$.
    If $C_{\text{avg}} \geq C_{\text{T}}$, update the clusters by including the phases of meter M.
    Otherwise, add meter M to the list of potential outliers L_O.
    Remove meter M from the list L_M.
  end while
end for

Fig. 1 Pseudocode of the clustering of smart meters.

Algorithm 2 Topology correction on the basis of dynamic data

Get the list of potential outliers L_O previously constructed.
while `length`$(L_O) \geq 0$ **do**
  Get an outlier O from the list L_O.
  Set the search area A, equivalent to a radius R around outlier O.
  Identify the list of feeders L_F with meters included in area A.
  while `length`$(L_F) \geq 0$ **do**
    Get a feeder F from the list L_F.
    Build a $3 \times n$ correlation matrix between the n phases of potential outlier O and the three reference phases of feeder F.
    Assign each phase of O to a cluster, in decreasing order of correlation C_n.
    Compute the average correlation of outlier O: $C_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} C_i$.
    Compute the correlation threshold $C_{\text{T}} = \max\ [C_{\max} - \alpha \cdot L - \beta \cdot N,\ C_{\min}]$.
    Generate a suggestion of topology correction if $C_{\text{avg}} \geq C_{\text{T}}$.
    Remove feeder F from the list L_F.
  end while
  Remove outlier O from the list L_O.
end while

Fig. 2 Pseudocode of the suggestions for topology correction.

4 Results

As previously hinted, the algorithmic solution presented has been tested on a real-world case involving two areas managed by RESA, a Belgian DSO. Naturally, the quality of dynamic data recorded by real smart meters varies significantly, with a significant number of missing or abnormal values. Such a situation surely presents an additional challenge, but it is representative of reality and enables to assess the robustness of the proposed methodology.

For this reason, the dynamic data go through some important pre-processing operations before being processed by the algorithmic solution presented in Section 3. Firstly, abnormal voltage values which are outside of a range [$V_{\min}$, $V_{\max}$] are replaced by *NaN* values. Secondly, the number of consecutive *NaN* values is determined, and time series with a gap larger than τ are discarded (typically one day). Finally, the remaining *NaN* values are replaced by the average voltage across all the time series present in the dataset for this specific area.

Before getting to the analysis of the results achieved, the hyperparameters used in the experiments are provided for the scope of reproducibility. These parameters are summarised in Table 1 hereafter.

Table 1 Hyperparameters used in the experiments presented.

Hyperparameter	Symbol	Value
Length of the voltage time series	T	20 days
Lower bound for the correlation threshold	$C_{\min}$	0.6
Upper bound for the correlation threshold	$C_{\max}$	0.8
Impact of the length of the power line on C_{T}	α	0.0001
Impact of the number of consumption points on C_{T}	β	0.005

As hinted in Section 3, the correlation is expected to decrease when the reference three-phase meter and the analysed meter are further apart. Such a behaviour not only happens when the length of the power line connecting the two meters increases and when the number of production or consumption points in between increases, but also when the structure of the power grid is more complex. Such a claim may seem perfectly logical or even trivial, but it is important to validate with practical data. Fig. 3 illustrates this phenomenon. In the following figures, the map is represented on the basis of GIS data. Each rhombus represents a smart meter, the colour indicating the result of the clustering operation (green for validation and red for outlier detection). The coloured circles protray traditional meters for which dynamic data are not available, their colour being indicative of their feeder. Finally, the bold lines represent the different power lines, coloured once again according to the associated feeder.

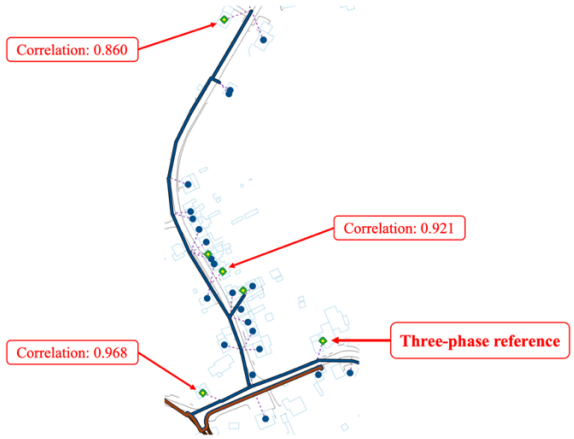

Fig. 3 Decrease in correlation when meters are further apart.

Very promising results are reported for the novel algorithmic solution presented in Section 3. Indeed, numerous ambiguous situations in the dataset of the DSO have been correctly identified by the algorithm, whose output was perfectly logical from a human's perspective. Moreover, the methodology proposed offers the key advantage of yielding interpretable and reasoned results, in contrast to black-box models. Such a feature is of utmost importance to DSOs, which have to manage critical infrastructures.

As a first example, Fig. 4 depicts a case for which it is almost certain that the algorithmic solution achieves a good result, despite having no ground truth available. In this figure, the outlier smart meter is represented by the red rhombus. The latter was initially assigned to the blue feeder on the right of the figure. However, the correlation with its reference (0.664) is below the threshold (0.71), indicating a potential outlier. The analysis of the neighbouring feeders reveals that there is a higher correlation (0.988) above the corresponding threshold with the reference of another feeder coloured in pale grey. Therefore, the suggestion to reallocate the smart meter analysed to this grey feeder is generated by the algorithmic solution. This seems perfectly logical from a human perspective, therefore validating this result.

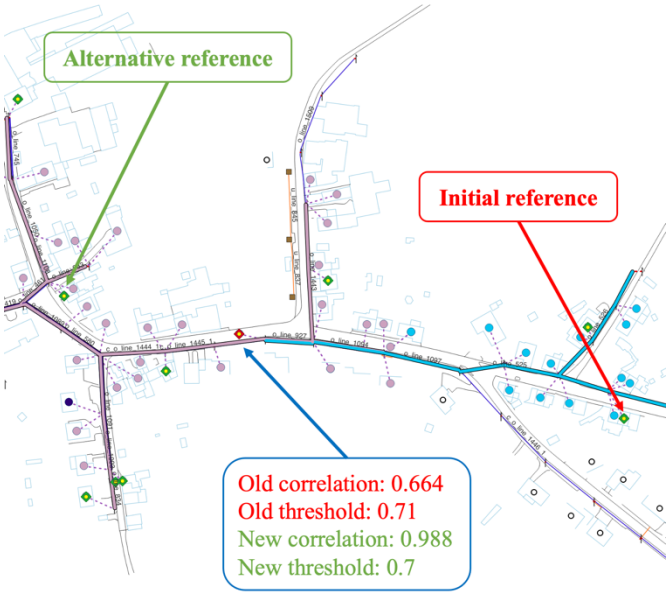

Fig. 4 Positive result achieved by the algorithmic solution.

As another example, Fig. 5 shows a more complicated case, for which the authors are confident about the output of the algorithmic solution in the absence of ground truth. In this configuration, the smart meter analysed is located near a crossroad, in a street presumably equipped with three power lines belonging to three different feeders. On the basis of the static data present in the database of the DSO, this smart meter is connected to the feeder coloured in yellow. Nevertheless, the analysis of dynamic data reveals that the correlation with the associated reference (0.717) is inferior to the threshold (0.75) despite having this reference being located very close. On the contrary, the correlation with the reference three-phase meter of the pink feeder (0.866) is superior to the corresponding threshold (0.771). Consequently, a suggestion is issue by the algorithmic solution, which appears to be logical from a human's perspective.

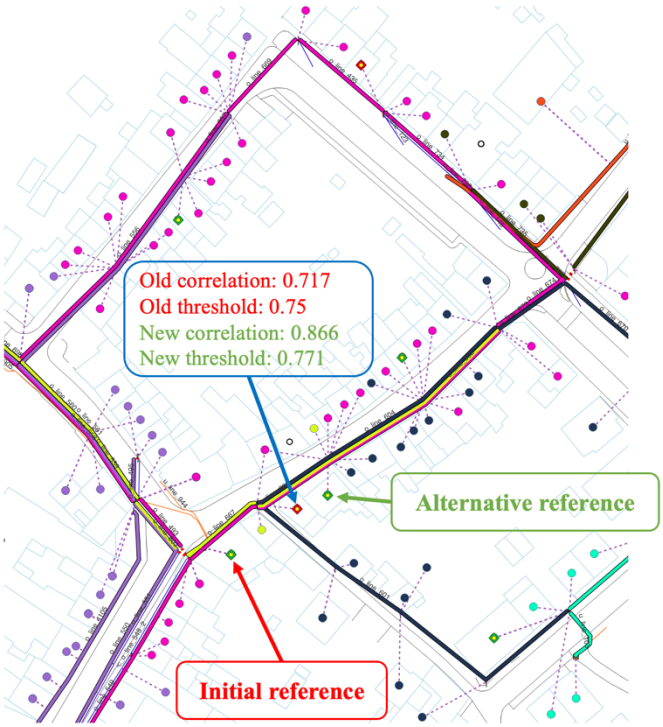

Fig. 5 Probable good detection of the algorithmic solution.

To finish with, the experiments conducted on this large-scale distribution network in practical situations have also revealed several other challenging scenarios for smart meter clustering. Typically, these inconsistencies occur in three configurations. Firstly, when there are very few smart meters available. Secondly, when the reference and the analysed smart meters are far apart. Lastly, when the structure of the power grid is more complex. An example of this observation is provided in Fig. 6 hereafter. In this case, the potential outlier detected is probably a false positive from a human's perspective. Indeed, there is a single feeder supposed in this street, and the other smart meters seem to be in line with the topological data. Such a problematic behaviour is assumed to originate from the specific situation of the smart meter:

- A division of the power line (Y shape) linking the reference to the analysed meter;
- The end of the power line (and municipality).

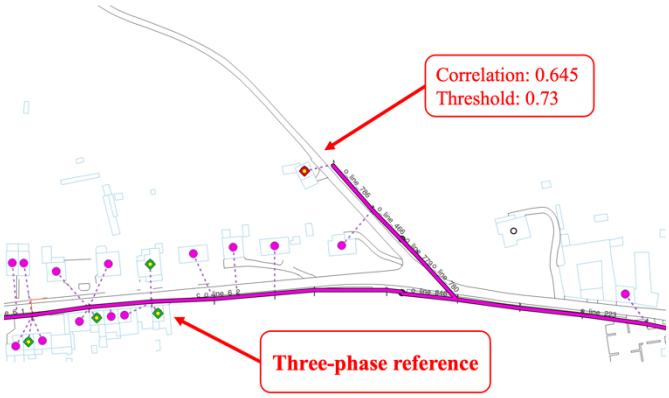

Fig. 6 Probable false positive detected by the methodology.

5 Future work

In addition to achieving interesting results, the algorithmic solution previously presented is expected to have room for further improvement. The following list summarises the main ideas to be investigated in the future:

- As previously hinted, the structure of the network (e.g., Y shape, end of line) could be taken into account when computing the correlation threshold.
- The order of the clustering operation could be altered so that the analysis starts from the reference selected and progressively propagates towards the end of the power line. The idea is to have a continuously updating reference, which would steadily grow and be composed of more than one three-phase smart meter or feeder. In this case, the reference time series would become a combination of the different signals with appropriate weights.
- In order to gain in robustness, the algorithmic solution presented could be applied on several time periods (e.g., the four seasons), followed by a comparison of the different results achieved.
- Once the clustering analysis is completed, the silhouette score of every sample within its cluster can be evaluated in order to identify additional outliers.
- If there is no three-phase element available (feeder or smart meter) to serve as reference, an alternative methodology is necessary to perform the clustering of the single-phase smart meters.

6 Conclusion

To conclude, this research work introduces a novel approach to verify the topology of the distribution power grid by taking advantage of the dynamic data collected by smart meters. More precisely, the methodology relies on the clustering of the electrical phase(s) measured by the smart meters, on the basis of correlation between voltage time series. The algorithmic solution presented yields positive results on a real-world case involving data collected by a Belgian DSO. These promising results inspired new ideas to be investigated in the future to further improve the performance of the algorithm. Moreover, the solution proposed is expected to serve other purposes, such as effectively balancing the electrical phases in the power grid.

7 References

[1] Ze-pu, G., Luo, Y., Ziwei, X., et al.: ‘Knowledge graph-based method for identifying topological structure of low-voltage distribution network’. The Journal of Engineering, 2020, 12, pp. 1177-1184

[2] Vassallo, M., Bahmanyar, A., Duchesne, L., et al.: ‘A systematic procedure for topological path identification with raw data transformation in electrical distribution networks’. 7th International Conference on Energy, Electrical and Power Engineering (CEEPE), Yangzhou, China, April 2024, pp. 707-715

[3] Seack, A., Kays, J., & Rehtanz, C.: ‘Generating low voltage grids on the basis of public available map data’. International Conference on Electricity Distribution (CIRED) Workshop, Rome, Italy, June 2014.

[4] Heidari-Akhijahani, A., Safdarian, A., & Aminifar, F.: ‘Phase identification of single-phase customers and PV panels via smart meter data’. IEEE Transactions on Smart Grid, 2021, 12, (5), pp. 4543-4552

[5] Olivier, F., Sutera, A., Geurts, P., et al.: ‘Phase identification of smart meters by clustering voltage measurements’. Power Systems Computation Conference (PSCC), Dublin, Ireland, June 2018, pp. 1–8