AUTOMATIC PHASE IDENTIFICATION OF SMART METER MEASUREMENT DATA

Frédéric OLIVIER
University of Liège – Belgium
frederic.olivier@ulg.ac.be

Damien ERNST
University of Liège – Belgium
dernst@ulg.ac.be

Raphaël FONTENEAU
University of Liège – Belgium
raphael.fonteneau@ulg.ac.be

ABSTRACT

This paper highlights the importance of the knowledge of the phase identification for the different measurement points inside a low-voltage distribution network. Besides considering existing solutions, we propose a novel method for identifying the phases of the measurement devices, based exclusively on voltage measurement correlation. It relies on graph theory and the notion of maximum spanning tree. It has been tested on a real Belgian LV network, first with simulated unbalanced voltage for which it managed to correctly identify the phases of all measurement points, second, on preliminary data from a real measurement campaign for which it shows encouraging results.

INTRODUCTION

PREMASOL is a project which aims a predicting, analysing and controlling photovoltaic production in low-voltage distribution networks. Within the framework of this project, a measurement campaign is ongoing in a sub-urban low-voltage (LV) distribution network in Belgium. Each house is equipped with a smart meter measuring the the voltages, the currents, the active and reactive power of the three phases. This measurement campaign has two goals: (i) The first goal is to validate the modelling that has been developed in a LV distribution system analysis tool. The active and reactive powers will be used as input and the resulting voltages will be compared to the measured ones. (ii) Another goal is for the distribution system operator (DSO) to gain insight on the three phase power flows inside its networks.

Since all measurements are three-phase, it is critical to be able to associate each of them to a physical phase of the network. However, this information may not be easily available depending on the technology used to transfer the data from the smart meters. Several solutions exist and will be discussed. However, none of them suited our requirements. As result, in this paper, we propose a methodology for automatically pairing the phases of any number of measurement devices with those of the network, solely using measurement data.

The paper is organized as follows. We first discuss the importance of being able to identify the phases of the smart meters. We then formalize the phase identification problem, shortly introduce the existing solutions, and describe our methodology. We finally provide an illustration of our method using both pseudo-measurements generated for the purpose of this study and real smart meter data.

ON THE IMPORTANCE OF PHASE INFORMATION

When a measurement device — smart meter or other — is placed in a low-voltage distribution network, it is critical to know to which phase of the network the three phase measurements correspond. This information is crucial for multiple reasons: In the absence of the latter, (i) phase measurements such as line to neutral voltage cannot be used to identify key features of the network. For example, it is impossible to display a voltage profile of the feeder for each phase. (ii) It is not possible to know from the measurements which phase of the network is the most loaded and how the DSO can better balance its network by changing the phases at customer place. This can help to increase the hosting capacity of a LV network with high penetration of photovoltaic (PV) panels [1] by mitigating the overvoltages that cause PV inverters to disconnect and induce a loss of earnings for the owner [2]. (iii) For research purposes, it is impossible to properly use the measurement to perform load flow on a model of the network and compare the resulting voltages.

THE PHASE IDENTIFICATION PROBLEM

For the purpose of explaining the phase identification problem, we will denote N_A, N_B and N_C the phases of the network and M_A^i, M_B^i and M_C^i the phases of the measurement device M^i. We already have established that the identity pairing N_A–M_A^i, N_B–M_B^i and N_C–M_C^i cannot be guaranteed because it depends on the connection of the cables of the smart meter and of the cables between the network and the house, an information not in the possession of the DSO of the network. The goal of the phase identification problem is, for each measurement point, to uniquely associate one phase of the measurements to one phase of the network. Since the reference is relative, the problem consists in clustering all the phase of the measurement points in three groups and then arbitrarily deciding which one corresponds to N_A, N_B or N_C.

EXISTING SOLUTIONS

Different approaches can be used to solve the phase identification problem.

Smart meters with PLC capabilities

To begin with, the measurements of the smart meter can be repatriated thanks to Power Line Communication (PLC). In such a case, the measurements are gathered at the secondary substation and PLC can be used to directly

identify the measured phases for each smart meter. The smart meter sends a different signal on each phase, a signal that is then used by the receiver to find the matching between it and the smart meter.

Phase identifiers

The second solution is to use a specific device which is based on GPS timing signals to compare the phase of an unknown voltage to a phase reference at the same instant of time. For this purpose, a base station is connected to a known phase, for example at the distribution transformer. A technician then proceeds to the phase identification of each house by reading the phase shift between the phase reference and the phase of the voltages at the smart meter. Examples of such phase identification devices are the "Phase ID 6000" by *Power Systems Integrity*, the "PVS 100i" by *Megger* or the "Phase Identifier" from *Orgo Corporation*. The drawbacks are their cost and the manpower required to perform the identification. Moreover, this operation requires access to smart meters located inside houses, which depends on customer availability.

MOTIVATIONS FOR AN ALTERNATIVE

For the present application, one would certainly advise in favour of the deployment of PLC-enabled smart meters to repatriate the data. However, there exists several cases where this is not a viable option: (i) The technical infrastructure is already up and running and does not use PLC (e.g. GPRS). (ii) The measurement devices are not smart meters but mobile measurement devices storing data in a local memory that has to be manually harvested, such as PQ boxes. (iii) The measurements to be analysed are from previous measurement campaigns and no record of the phases has been included.

This highlights the need for an alternative. Two directions oppose each other: one based on equipment and man power, the other based on the analysis of the measurements themselves, which is the topic of this paper. Of course, the two are not mutually exclusive and the solution chosen by DSOs will most likely be based on their financial appeal. In that regard, the proposed method should be of interest as it can be significantly cheaper.

This situation was observed within the PREMASOL project. One of the partners specialises in photovoltaic monitoring and already possesses an infrastructure to repatriate data through GPRS so this option was favoured over PLC. Moreover, the DSO wanted a less expensive solution than to invest in phase identification equipment and dedicate manpower to this task.

In this paper, we propose an alternative based on the measurements and the unbalanced nature of the power flow inside the networks, to cluster the measurements by phase using correlation. Although time-series clustering is a well-studied problem [3], it has, to the best of our knowledge, never been applied to the phase identification problem.

ASSUMPTIONS

The methodology is designed for unbalanced three-phase low-voltage distribution networks where houses have a smart meter – and/or where mobile measurement devices are placed – measuring phase-to-neutral voltages for each phase. We also assume that the measurements are synchronized and have the same sampling period of at least one minute to be able to capture the voltages variations. As explained in [4], "averaging data over periods longer than a minute is shown to under-estimate the proportions of both [electricity] export and import."

THE IDENTIFICATION ALGORITHM

The key idea on which the identification algorithm relies is that LV distribution networks are intrinsically unbalanced. Indeed, house appliances are mainly single phase. This unbalanced load creates an unbalance in voltages which can be used to cluster the measurements, phase by phase.

The algorithm relies on graph theory, more specifically on the notion of a maximum spanning tree (MST), to select the pairings that will maximize the correlation between the voltages of the nodes. Using graph theory and maximum spanning trees comes naturally since any electrical network can be viewed as graph where nodes are buses and branch electrical lines, and distribution networks are usually operated in a radial fashion, so they can be viewed as trees.

The different steps of the algorithm are the selection of the data, the creation of a complete graph where the branches are weighted by a correlation coefficient, and the selection of the most relevant branches by a maximizing spanning tree algorithm. The algorithm ends with the clustering of the measurements into three groups.

The next subsections discuss these steps in more detail.

Selecting a time window for the data

First, a time window is selected; for example, one day, and to each measurement point three voltage time series are associated, one for each phase-to-neutral voltage.

Creating the graph

Let us define a pairing between a measurement M^i and a measurement M^j as a set of two mappings: a mapping from the three phases of M^i to the three phases of M^j, and an associated mapping from the three phases of M^j to three phases of M^i. Given that there are three phases, there are at most $3! = 6$ different sets of pairings, detailed hereafter.

Pairing 1	$M_A^i - M_A^j$	$M_B^i - M_B^j$	$M_C^i - M_C^j$
Pairing 2	$M_A^i - M_A^j$	$M_B^i - M_C^j$	$M_C^i - M_B^j$
Pairing 3	$M_A^i - M_B^j$	$M_B^i - M_A^j$	$M_C^i - M_C^j$

Pairing 4	$M_A^i - M_B^j$	$M_B^i - M_C^j$	$M_C^i - M_A^j$
Pairing 5	$M_A^i - M_C^j$	$M_B^i - M_A^j$	$M_C^i - M_B^j$
Pairing 6	$M_A^i - M_C^j$	$M_B^i - M_B^j$	$M_C^i - M_A^j$

At this point, it can be noted that pairings 1 to 3, and 6, are symmetrical in the sense that the two associated mappings are identical. This is not the case for pairings 4 and 5: pairing 4 becomes pairing 5 and *vice versa.*

The next step of the algorithm is the creation of a complete graph where each node is connected to all the others. It means that, for a graph with n nodes, there are $n(n-1)/2$ branches. For each branch and its two end nodes, there are 6 different possible pairings as defined above.

For each pairing, the sum of the correlation coefficients between the three pairs of voltage time series is computed. For example, the correlation coefficient for the first and second pairing are computed hereafter:

$$\rho_1^{i,j} = corr\left(M_A^i, M_A^j\right) + corr\left(M_B^i, M_B^j\right) + corr\left(M_C^i, M_C^j\right)$$
$$\rho_2^{i,j} = corr\left(M_A^i, M_A^j\right) + corr\left(M_B^i, M_C^j\right) + corr\left(M_C^i, M_B^j\right)$$

The same formula can be applied for the other pairings.

Pearson's measure is used is assess the correlation between the time series:

$$corr(X,Y) = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{T}(X_t - \bar{X})^2}\sqrt{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}}$$

where X and Y are to time series of length T with a mean value of $\bar{X}$ and $\bar{Y}$.

Each branch of the graph represents the pairing that results in the maximum correlation coefficient, the weight of the branch being this maximum coefficient.

$$w_{i,j} = \max\left(\rho_1^{i,j}, \rho_2^{i,j}, \rho_3^{i,j}, \rho_4^{i,j}, \rho_5^{i,j}, \rho_6^{i,j}\right)$$

When storing the information on the pairing that maximizes correlation between two nodes, it is important to have a data structure that can differentiate between the two directions the branch can be traversed, because, as explained above, pairing 4 in one direction is pairing 5 in the other.

Finding the maximum spanning tree

Next, the Prim algorithm [5] is used to find the maximum spanning tree of the complete graph. It selects the edges that will bring the maximum total correlation between the nodes. The result is a tree where each branch represents the pairing which must be used to link measurements from the parent node to the child node. By using a tree, we ensure that there is no cycle inside the network and that there is only one possible succession of pairings from one node to any other.

Selecting the reference and clustering the phases

Once the tree is computed, we know how the phases of two adjacent nodes are paired, but this information is relative to the phases of the parent. The final step is to select a reference node to start from, and to traverse the entire tree structure, from parent to children, applying the pairing of each branch to uniquely select which cluster each phase measurements belongs to. The final results of the algorithm are three sets, $\mathcal{C}_A$, $\mathcal{C}_B$ and $\mathcal{C}_C$ where, for example,

$$\mathcal{C}_A = \{M_A^1, M_B^2, M_B^3, \dots\}$$
$$\mathcal{C}_B = \{M_B^1, M_A^2, M_C^3, \dots\}$$
$$\mathcal{C}_C = \{M_C^1, M_C^2, M_A^3, \dots\}$$

that can be arbitrarily associated to the phases of the network N_A, N_B and N_C.

Including information on the topology of the distribution network

If more information on the structure of the electrical network is available, it can be used to reduce the number of branches from the complete graph. If nodes are far apart and on different feeders, the branch linking them can be discarded, reducing the number of edges and thus simplifying the resolution of the maximum spanning tree problem. One extreme option could be to directly define the tree by linking the nodes that are closest to each other, thus eliminating the need for the maximum spanning tree step. However, it is obviously not recommended because this eliminates a powerful step of the algorithm which can restructure the network based on correlation. For example, if measurement data were corrupted or if one house was associated with the wrong feeder, the algorithm will be able to circumvent those errors. Finally, it would be a strong shortcut to assume that the closer the measurement points are geographically, the stronger the correlation between the voltages as voltage variations are mainly due to the line impedance. So, we advise the reader to suppress branches with parsimony and only those that are without doubt irrelevant.

TEST NETWORK

The test network used for this study is an existing Belgian low-voltage distribution network, composed of three feeders made with underground cables of the type EVAVB-F2 3x95 + 1x50. It is located in a suburban area and has been modelled according to [6] based on the data provided by the DSO (topology, line length, cable type, etc.). Detailed unbalanced three-phase four-wire modelling of the network has been used.

Model for the dwellings: load and photovoltaic units

The network is composed of 32 houses, all of which have a three-phase 400/230 V connection of various length with a cable of type EXAVB 4x10. Five of these houses are equipped with photovoltaic units.

The energy consumption of the house is modelled using consumption profiles created with [7]. Several alterations have been made to the code created by Widén and Wäckelgård in order to allow the creation of unbalanced load profiles.

First, the appliances have been classified as single phase or three phase. Each single-phase appliance has been allocated to one of three groups based on good practice,

trying to balance the load in each group as optimally as possible. Each time the profile generator is run, appliances are clustered in the same groups, however, the clusters are randomly allocated to a specific phase. At this point, rather than calculate the sum of the consumption of all appliances, appliance consumptions are summed phase by phase, adding one third of the three-phase appliances. An example of a load profile can be seen in Figure 1.

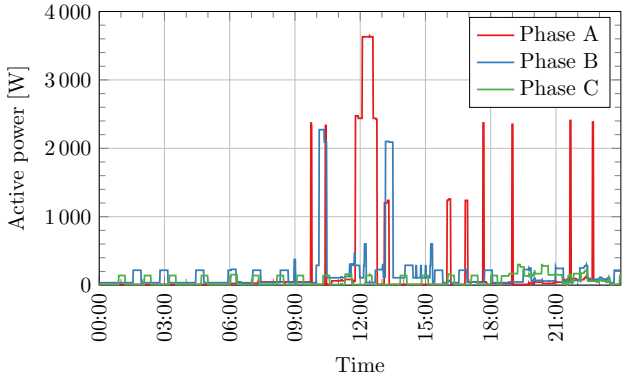

Figure 1 – Example of an unbalanced load profile.

The production of the photovoltaic panels is based on the production of a typical photovoltaic unit in Belgium, scaled with respect to the peak power of each unit. The consumption and production profile have been generated for an arbitrarily chosen day: Thursday, the 5th of May, 2016.

Model for the medium voltage network

The medium voltage network is modelled as a Thevenin equivalent. The phase-to-phase voltage of the equivalent is fixed at 420 V and the impedance at $0.0059 + j \cdot 0.0094\ \Omega$.

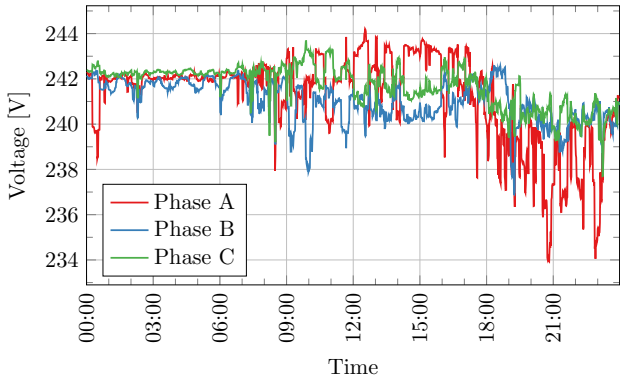

Figure 2 – Voltages for measurement point M^{25}.

Pseudo-measurement generation

An unbalanced load flow algorithm based on [6] and implemented in Python is used to compute the currents in the lines and the voltages at each node with a resolution of 1 minute to capture the variability of the loads. The three phase-to-neutral voltages at the bus where a house is connected are exported so that they can be used in the next step. Such voltages are displayed in Figure 2 for measurement point M^{25}.

RESULTS

The algorithm was implemented in Matlab. To compute the maximum spanning tree, the Prim algorithm (minimum spanning tree) is used with the opposite of the branches' weight.

With pseudo-measurements

Because they are generated using a LV network simulator, phase measurements are already sorted, so the first step is to randomly permute them for each measurement point. Then, the identification algorithm is run on the data.

The most important result is that the algorithm successfully manages to identify all the phases and to cluster them the proper way, regardless of the initial permutation. As an illustration, Figure 3 shows the voltages that were clustered by the algorithm into one single group.

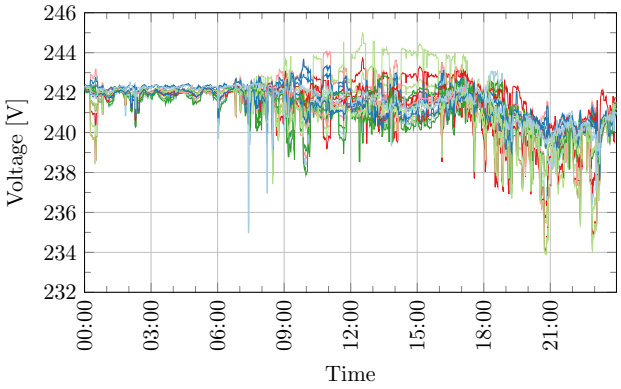

Figure 3 – Example of a cluster of time series after the algorithm.

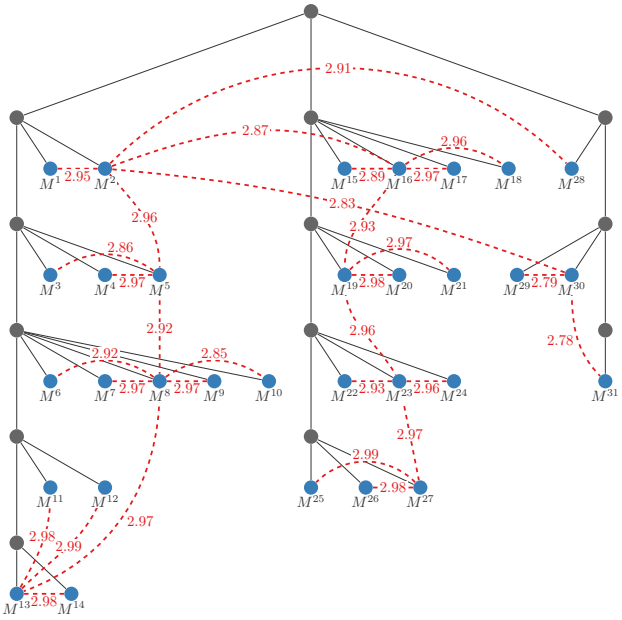

Figure 4 – The topology of the electrical network is in grey. The upper node corresponds to the distribution transformer and the nodes that are numbered are measurement points (blue). The maximum spanning tree is in red (dashed) and the weight of the edges is the maximum correlation coefficient.

The maximum spanning tree output by the algorithm is displayed in Figure 4, superimposed on the structure of the electrical network. It can be seen that the structure of the MST is coherent with the structure of the network as no nodes are connected between different feeders except at the root of the network. The edges are weighted by the correlation coefficient of the correlation maximizing pairing. It can be seen that all values are close to 3, the maximum, indicating an excellent correlation between the nodes.

With real measurements

At the time of writing this paper, the roll-out of the smart meters for the purpose of the measurement campaign is still ongoing and all houses have not yet been equipped. Regardless, the algorithm has been applied to the measurements that have already been collected. The resulting tree is displayed in Figure 5.

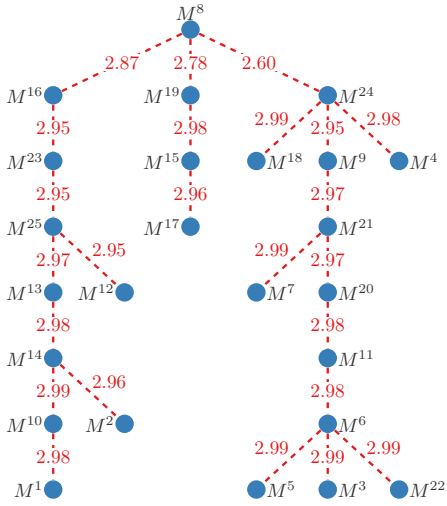

Figure 5 – Maximal spanning tree for the real measurement data where the weight of the edges is the maximum correlation coefficient.

It can be seen that the majority of edges have a strong correlation coefficient, except the one linking nodes 8 and 24. The reason is certainly that the two measurement points are too far apart, either due to the lack of a smart meter in between, or due to a large impedance between them. It is information that can be further investigated by the DSO in its phase identification process. In any case, we advise the installation of a smart meter at the low-voltage side of the distribution transformer to provide a measurement point that can be used to link different feeders.

Finally, the method can be used to check if a measurement point's location was incorrect. It could be observed when a branch of the MST is not coherent with the topology of the network, and has a low correlation coefficient.

CONCLUSION

We have proposed a phase identification algorithm which performs exactly as planned when applied to measurements purposely generated. In addition to phase identification, the algorithm outputs a maximum spanning tree which provides insight into the structure of the electrical network and the measurements that are more correlated. Furthermore, the first results from real measurements are extremely encouraging as the correlation coefficients are close to their maximum.

Several research questions arise from this study. It would be interesting to investigate the behaviour of the algorithm when the sampling period of the measurement is increased (from 1 to 10 minutes for example). This should lead to a decrease in the imbalance of the voltages and lead to a decrease of the overall correlation coefficients. The next steps are to further analyse the real measurement data in regards to their location in the network, and to provide an explanation and/or a solution to the links of the graph which have a poor correlation coefficient.

ACKNOWLEDGEMENTS

The authors thank the Walloon Region for the financial support of the PREMASOL project, GreenWatch and Réseau d'Energies de Wavre for fruitful discussions and data.

REFERENCES

[1] R. A. Walling, R. Saint, R. C. Dugan, J. Burke, and L. A. Kojovic, "Summary of Distributed Resources Impact on Power Delivery Systems," *IEEE Trans. Power Deliv.*, vol. 23, no. 3, pp. 1636–1644, Jul. 2008.

[2] F. Olivier, P. Aristidou, D. Ernst, and T. Van Cutsem, "Active Management of Low-Voltage Networks for Mitigating Overvoltages Due to Photovoltaic Units," *IEEE Trans. Smart Grid*, 2016.

[3] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[4] A. Wright and S. Firth, "The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations," *Appl. Energy*, vol. 84, no. 4, pp. 389–403, 2007.

[5] R. C. Prim, "Shortest Connection Networks And Some Generalizations," *Bell Syst. Tech. J.*, vol. 36, no. 6, pp. 1389–1401, Nov. 1957.

[6] R. M. Ciric, A. P. Feltrin, and L. F. Ochoa, "Power flow in four-wire distribution networks-general approach," *IEEE Trans. Power Syst.*, vol. 18, no. 4, pp. 1283–1290, Nov. 2003.

[7] J. Widén and E. Wäckelgård, "A high-resolution stochastic model of domestic activity patterns and electricity demand," *Appl. Energy*, vol. 87, no. 6, pp. 1880–1892, Jun. 2010.