Phase Identification of Smart Meters by Clustering Voltage Measurements

Frédéric OLIVIER, Antonio SUTERA, Pierre GEURTS, Raphaël FONTENEAU, Damien ERNST
Department of Electrical Engineering and Computer Science
University of Liège, Belgium
{frederic.olivier, a.sutera, p.geurts, raphael.fonteneau, dernst}@uliege.be

***Abstract*—When a smart meter, be it single-phase or three-phase, is connected to a three-phase network, the phase(s) to which it is connected is (are) initially not known. This means that each of its measurements is not uniquely associated with a phase of the distribution network. This phase information is important because it can be used by Distribution System Operators to take actions in order to have a network that is more balanced.**

In this work, the correlation between the voltage measurements of the smart meters is used to identify the phases. To do so, the constrained k-means clustering method is first introduced as a reference, as it has been previously used for phase identification. A novel, automatic and effective method is then proposed to overcome the main drawback of the constrained k-means clustering, and improve the quality of the clustering. Indeed, it takes into account the underlying structure of the low-voltage distribution networks beneath the voltage measurements without a priori knowledge on the topology of the network. Both methods are analysed with real measurements from a distribution network in Belgium. The proposed algorithm shows superior performance in different settings, e.g. when the ratio of single-phase over three-phase meters in the network is high, when the period over which the voltages are averaged is longer than one minute, etc.

***Index Terms*—Voltage correlation, clustering, constrained k-means clustering, phase identification, smart meter, three-phase distribution network.**

I. INTRODUCTION

When a smart meter, be it single-phase or three-phase, is connected to a three-phase network, the phase(s) to which it is connected is (are) initially not known. This means that each of its measurements is not uniquely associated with a phase of the distribution network. This can be achieved by solving the phase identification problem, i.e. associating the phases of the network to those of the smart meter. It is equivalent to clustering the measurements in three groups, one for each phase of the network. This phase information is important because it can be used by Distribution System Operators (DSO) to take actions in order to have a network that is more balanced.

In this work, the correlation between the voltage measurements of the smart meters is used to solve the phase identification problem. To do so, the constrained k-means clustering method is first introduced. It will be used as a reference for the performance as it has been previously used for phase identification [1]. A novel, automatic and effective method is then proposed to overcome the main drawback of the constrained k-means clustering and improve the quality of the clustering. Indeed, it takes into account the underlying structure of the low-voltage (LV) distribution networks beneath the voltage measurements without a priori knowledge on the topology of the network. Both methods are analysed with real measurements from a distribution network in Belgium, in order to give insight on how to parametrise them, and generalise their performance in different settings, by varying the ratios of single-phase over three-phase meters in the network, increasing the period over which the voltages are averaged, etc.

The paper is organised as follows: the phase identification problem is motivated in the first section and formalised in the second. In section 3, the two clustering algorithms are presented. We then describe the LV distribution network used for the computations and the assessment of the algorithms. Sections 4 and 5 present the parametrisation of the algorithms and their performance in different settings. Finally, section 6 summarises the advantages and pitfalls of the proposed clustering methods to identify the phase of smart meters in a LV distribution network.

II. ON THE IMPORTANCE OF PHASE IDENTIFICATION

LV distribution networks are intrinsically unbalanced due to the currents that are consumed by single-phase household appliances, or produced by single-phase distributed generation units. This imbalance reduces the hosting capacity of the network, i.e. the maximum amount of distributed generation that can be connected without violating the operational constraints in the network [2]–[5]. Thus, the first action to increase the hosting capacity is to better balance production on the three phases of the network. The phase information allows the distribution system operators (DSOs) to do so by changing the phase(s) to which their customers are connected.

Moreover, three-phase power flow simulations take as input the active and reactive powers that are consumed or produced at each connection point, phase by phase. To obtain simulations accurately reflecting real operations, the phase information is required to know to which phase of the network the active and reactive power measurements correspond. This

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

also allows for the comparison between the voltages that are measured and voltages that are simulated.

III. The phase identification problem

In this paper, we consider that the meters can either be single phase, i.e. connected between a phase and the neutral, or three phase, i.e. connected to the three phases and the neutral. This implies that the network to which the meters are connected is a four-wire, three-phase and neutral star-shaped distribution network. Let

$$\mathcal{S} = \{S_j\}, j \in \mathcal{J} = 1...J$$

be the set of all meters, where J is the total number of smart meters. The meters measure several electrical quantities that can be aggregated, such the total active power that is consumed, or phase related, such as phase-to-neutral voltages. The latter measurements are identified by the letters r_j, s_j, t_j, among which each corresponds to one phase of the meter. When a smart meter is single phase, only the index r_j is used.

As explained in the introduction, there is no coherence between r, s and t at the network level. They are only valid at the meter level. If the phases of the network are labelled a, b, c, the goal of the phase identification is thus to associate the phase r_j, s_j and t_j of each smart meter, S_j, to one phase of the distribution network.

It is equivalent to clustering the phases of the meters in three groups, which will be denoted $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, and then associating each group to one phase of the network by taking a common reference. Moreover, the physical nature of the problem imposes constraints on the clustering process. Since meter phases are electrically connected to different phases of the network, two indices linked to the same smart meter cannot be placed in the same cluster, i.e. r_j, s_j, t_j must be placed in different clusters. Of course, single-phase meters do not impose such constraints and the phase index can be placed in a cluster without restrictions.

IV. Existing solutions

Two types of solution oppose each other to solve the phase identification problem: *manual* ones, which require a technician to proceed, manually, to the phase identification and *automatic* ones which either use a built-in function of the smart meters or perform an analysis of the measurements.

A. Manual methods

On the one hand, phase identifiers are equipment usually composed of two parts, with one of them connected at a reference point in the network and the other one used by a technician to identify the phases at the customers' premises. A primary technology uses clocks that are synchronised with GPS, to compare the voltage angle between the reference voltage and the voltage measured by the technician. A second technology broadcasts different signals in the three phase conductors, and the technician proceeds to the phase identification by reading the signal in each of the smart meter's conductors. The main drawbacks are the equipment cost and the need to send a technician to the premises. Moreover, these methods may also be prone to human error.

B. Automatic methods

On the other hand, recent smart meters, which uses Power Line Carrier (PLC) technology to transfer the measurements, have a built-in function to identify the phases. However, if it does not have this function or uses a technology other than PLC, e.g. General Packet Radio Service (GPRS), a remote solution would instead be to use the measurements collected by the smart meter to perform the phase identification.

In the literature, it has been proposed to use energy measurements [6] or voltage measurements [7], [8] to (re)-discover the network topology, including the phase identification of the smart meters. Paper [6] proposes a method using graph theory based on a principal component analysis (PCA), whereas [7], [8] use a correlation-based approach exploiting the similarities between the voltage measurements of the same phase (k-means clustering). Papers [1], [9], [10] explicitly identify the phases with either energy or voltage measurements. [9] applies graph theory and PCA to identify the phases of a simulated dataset, taking into account the effect of noise. [10] uses correlation to find how to locally associate the phases of each smart meter to the phases of a reference meter. Finally, instead of directly employing correlation, [1] extracts features on which they perform a k-means clustering.

This paper extends the work done in [11]. Its contributions are four-fold: (i) It proposes a novel algorithm with a global network, using the advantages of both graph theory and correlation to identify the measurements that should be linked together and cluster them. (ii) The performance of the algorithm is assessed in comparison to those of a constrained k-means clustering performed on the voltage measurements. (iii) Unlike previous references, the algorithms are designed for the specificities of European LV distribution networks. (iv) The algorithms are tested on real measurements from a distribution network in Belgium, in a variety of settings.

V. Description of the proposed methods

After a description of the distribution network characteristics that provides an insight on how to measure similarity between two voltage measurements, the constrained k-means and the novel approach are presented. It should be mentioned that both methods are automatic and do not require additional equipment. Moreover, they are solely based on voltage measurements and do not require a priori knowledge of the network topology, such as the partition of smart meters between feeders. This point is important since this information is often not reliable due to the various reconfigurations of LV distribution networks.

A. Definition of the distance between two voltage measurements

Correlation-based approaches seem very well adapted to the phase identification problem and have been successfully used in [8], [10]. The proposed algorithms also use Pearson's correlation to assess the similarities between two time series

M_1 and M_2, gathering the measurements of the phase-to-neutral voltage magnitudes:

$$PC(M_1,M_2)=\frac{\sum_{\tau=1}^{n}\left(M_{1,\tau}-\hat{M}_1\right)\left(M_{2,\tau}-\hat{M}_2\right)}{\sqrt{\sum_{\tau=1}^{n}\left(M_{1,\tau}-\hat{M}_1\right)^2}\sqrt{\sum_{\tau=1}^{n}\left(M_{2,\tau}-\hat{M}_2\right)^2}}$$

where $\hat{M}_1$ and $\hat{M}_2$ are the mean values of M_1 and M_2, and n the length of both time series. Indeed, the voltage at the end of an electrical line is equal to the sum of the voltage at its beginning and the voltage drop caused by the current flowing through the impedance. Thus, points that are separated by a line with a small impedance, or which are traversed by a small current, have voltages that are more correlated. Since phases are not electrically connected to each other, voltage measurement from different phases are less correlated. Other similarity metrics could be of interest, such as the cosine similarity, if the same approach were to be adapted to other applications with a different type of data.

Finally, let the set of voltage measurements be $\mathcal{M}=\{M_i\}$, with $i\in\mathcal{I}=\{r_1,s_1,t_1,r_2,...,r_J,s_J,t_J\}$, where M_i is a voltage time series and i is its index in the total set of time series, not the time index. The distance between two voltage measurements is defined as

$$d(M_l,M_i)=1-PC(M_l,M_i),\quad \forall l,i\in\mathcal{I}$$

so that the distance is equal to 0 if the measurements are perfectly correlated.

B. Reference algorithm: Constrained k-means Clustering

k-means clustering [12] aims at partitioning a set of observations into k clusters. At each iteration, the observations are associated to the cluster with the *closest* centre, in the sense of a distance metric (e.g., Euclidian distance, correlation measure, etc.) that assesses the similarity between an observation and a centre. The centres are usually computed by averaging the observations already associated to the cluster. In the phase identification case, the number of clusters, k, is equal to the number of phases, i.e. $k=3$. However, this algorithm does not take into account the constraints that the measurements from a same three-phase smart meter must be placed in different clusters. The k-means clustering algorithm is thus modified to correspond to the one proposed in [13], which introduces background knowledge through constraints..

Formally, the constrained k-means (CKM) clustering algorithm for phase identification (as used in references [1], [13], [14], detailed in Algorithm 1 and illustrated in Figure 6) works as follows: Three empty clusters are defined based on the number of phases. The initial centres are equal to the measurements of the root three-phase smart meter given as input. This ensures that initial centres are measurements from three distinct phases. At each iteration, all smart meters are examined: each voltage time series of the smart meter is associated with the cluster whose center is the closest, in such a way that only one measurement can belong to a cluster to satisfy the three-phase constraint. For the sake of implementation, the pair measurement-cluster with the smallest distance is first selected, then the second pair (excluding the measurement and cluster from the first pair), and then the last pair is trivially associated to the remaining cluster. For a single-phase element, the measurement is directly associated to the closest centre. Once all smart meters have been examined, centres are updated by averaging all the measurements that have been assigned to them. If those updates are still above the given convergence threshold, or if the maximal number of iterations is not reached, then the process keeps iterating.

Algorithm 1 Constrained k-means Algorithm

```
Inputs:
  1) Set of measurements $\mathcal{M}$
  2) Set of smart meters $\mathcal{S}$
  3) Three-phase root smart meter $S_{j_0}$
  4) Distance metric $d(\cdot,\cdot)$
Output:
  • Three clusters satisfying the smart meter constraints: $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$.
Algorithm:
  1) Let $c_1=M_{r_{j_0}}, c_2=M_{s_{j_0}}, c_3=M_{t_{j_0}}$ be the initial cluster centres
     of $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$.
  2) Let $\mathcal{J}_1$ be the subset of the single-phase smart meter indices and $\mathcal{J}_3$
     the subset of three-phase smart meter indices.
  3) $\mathcal{C}_k=\emptyset$ with $k=1,2,3$.
  4) $\forall S_j$, with $j\in\mathcal{J}_1$,
        $k^*=\arg\min_{k=\{1,2,3\}} d(c_k,M_{r_j})$
        $\mathcal{C}_{k^*}=\mathcal{C}_{k^*}\cup r_j$
  5) $\forall S_j$ with $j\in\mathcal{J}_3$,
     a) $\{k_1^*,i_1^*\}=\arg\min_{k=\{1,2,3\},\ i=\{r_j,s_j,t_j\}} d(c_k,M_i)$
        $\mathcal{C}_{k_1^*}=\mathcal{C}_{k_1^*}\cup i_1^*$
     b) $\{k_2^*,i_2^*\}=\arg\min_{k=\{1,2,3\}\backslash k_1^*,\ i=\{r_j,s_j,t_j\}\backslash i_1^*} d(\mathcal{C}_k,M_i)$
        $\mathcal{C}_{k_2^*}=\mathcal{C}_{k_2^*}\cup i_2^*$
     c) $k_3^*=\{1,2,3\}\backslash\{k_1^*,k_2^*\},\ i_3^*=\{r_j,s_j,t_j\}\backslash\{i_1^*,i_2^*\}$
        $\mathcal{C}_{k_3^*}=\mathcal{C}_{k_3^*}\cup i_3^*$
  6) For each cluster $\mathcal{C}_k$, update its center by averaging all the measure-
     ments $M_i$, $\forall i\in\mathcal{C}_k$.
  7) Iterate between (3) and (6) until the algorithm reaches convergence
     (i.e., changes in centres are smaller than a given threshold $\epsilon$) or the
     maximal number of iterations.
```

A drawback of the k-means clustering approach is the fact that, between two subsequent iterations, measurements associated with a given cluster are averaged to compute the cluster centre for the next iteration. Averaging measurements may destroy some information by eliminating small variations in the measurements.

C. Proposed algorithm: Constrained Multi-tree Clustering

This paragraph explains the motivation for a tree-structured algorithm to cluster the phase measurements. Distribution systems are usually operated radially, i.e. there are no electrical loops, with one point of connection between the MV network and the LV network (the distribution transformer). From a graph theory perspective, the network can be seen as a tree, where the distribution transformer is the root and the connection points are the leaves. Indeed, a tree is defined as an undirected graph in which any two nodes are connected by exactly one path, in other words without cycles. It makes sense to use a tree structure to cluster the measurements, as explained in [11], because the tree structure of the network creates an underlying structure between the voltage measurements. Indeed, the voltages at two points that are neighbours in the network tree are more correlated.

The Constrained Multi-tree (CMT) algorithm was inspired by Prim's algorithm, which is used to calculate minimum spanning trees in weighed graphs. Prim's algorithm starts at a root node and makes the tree grow gradually by adding the node whose branch will add the lowest weight to the tree. Prim's algorithm cannot be applied in this form to cluster the measurements because its output is only one tree. The algorithm is modified to output three trees, hence three clusters.

Let

$$\Delta(\mathcal{C}_k, M_i) = \min_{l \in \mathcal{C}_k} d(M_l, M_i)$$

be the definition of the distance between a cluster $\mathcal{C}_k$ and a measurement M_i. Initially, the clusters contain only one measurement, those of the root smart meter given as input to the algorithm. For the same reason as for the previous algorithm, the root smart meter needs to be three-phase. Measurements will be sequentially added to the root measurements to form the trees in the cluster, as explained in Algorithm 2 and illustrated in Figure 7. At each iteration, a distance δ_i is associated to the measurement M_i from a smart meter, that is the closest to a cluster, while satisfying the constraints. A potential cluster κ_i and a potential predecessor π_i are also associated, corresponding to the cluster and its measurement, which minimises the distance. The constraints are taken into account by selecting the best pair between the remaining measurements for a smart meter and the remaining clusters. If the measurements of a smart meter are already in a cluster, the distance, potential cluster, and potential predecessor are computed without considering those measurements and corresponding clusters.

Once the distances have been computed, the measurement, which is not yet in a cluster and whose distance is the smallest, is added to its potential cluster, and edge is created with its potential predecessor. By doing so, the trees grow at each iteration by adding the measurement with the minimum cost, i.e. whose connection to the minimum spanning tree will make the sum of the weight of the edges increase by the least amount. The process is repeated until all measurements are clustered.

VI. Test system

A. The low-voltage distribution network

The algorithm is tested on voltage measurements from a Belgian LV distribution network, which is composed of five feeders with a star configuration 400V/230V.

Algorithm 2 Constrained Multi-Tree Algorithm

```
Inputs:
  1) Set of measurements $\mathcal{M}$
  2) Set of smart meters $\mathcal{S}$
  3) Three-phase root smart meter $S_{j_0}$
  4) Distance metric $d(\cdot,\cdot)$
Output:
  • Three clusters satisfying the smart meter constraints: $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$.
Algorithm:
  1) Let $\mathcal{C}_1 = \{r_{j_0}\}, \mathcal{C}_2 = \{s_{j_0}\}, \mathcal{C}_3 = \{t_{j_0}\}$ be the initial clusters,
     where $r_{j_0}, s_{j_0}, t_{j_0}$ are the phase indices of the root smart meter $S_{j_0}$.
  2) Let $\mathcal{J}_1$ be the subset of single-phase smart meters and $\mathcal{J}_3$ the subset
     of three-phase smart meters.
  3) $\delta_i = +\infty, \forall i \in \mathcal{I}$
  4) $\forall S_j$ with $j \in \mathcal{J}_1$, whose measurement index is not yet in a cluster,
       a) $k^* = \arg\min_{k=\{1,2,3\}} \Delta(\mathcal{C}_k, M_{r_j})$
       b) $\delta_{r_j} = \Delta(\mathcal{C}_{k^*}, M_{r_j})$
          $\pi_{r_j} = \arg\min_{l\in\mathcal{C}_{k^*}} d(M_l, M_{r_j})$
          $\kappa_{r_j} = k^*$
  5) $\forall S_j$, with $j \in \mathcal{J}_3$, whose three measurements are not yet in a cluster,
       • Let $\mathcal{I}^* = \{r_j, s_j, t_j\} \cap (\cup_{k=1}^3 \mathcal{C}_k)$ be the set of measurement
         indices from the smart meter $S_j$ that are already in a cluster, and
         the set of corresponding clusters $\mathcal{K}^*$.
       • Then,
           $\{k^*, i^*\} = \arg\min_{\substack{k = \{1,2,3\}\setminus\mathcal{K}^* \\ i = \{r_j, s_j, t_j\}\setminus\mathcal{I}^*}} \Delta(\mathcal{C}_k, M_i)$
           $\delta_{i^*} = \Delta(\mathcal{C}_{k^*}, M_{i^*})$
           $\pi_{i^*} = \arg\min_{l\in\mathcal{C}_{k^*}} d(M_l, M_{i^*})$
           $\kappa_{i^*} = k^*$
  6) The measurement index with the smallest distance $\delta$ is added to the
     corresponding cluster $\kappa$:
           $i^* = \arg\min_{i\in\mathcal{I}\setminus\cup_{k=1}^3 C_k} \delta_i$
           $\mathcal{C}_{\kappa_{i^*}} = \mathcal{C}_{\kappa_{i^*}} \cup i^*$
     Create an edge between $i^*$ and $\pi_{i^*}$.
  7) Iterate between (3) and (6) until all measurements are assigned to a
     cluster.
```

The network supplies 89 houses, among which 74 are equipped with a three-phase smart meter and two with a single-phase smart meter. Three houses are equipped with both a regular smart meter and a night-exclusive one, the latter not providing voltage measurements during the day from 7:00 to 22:00.

Each smart meter provides phase-to-neutral voltage measurements as a one-minute average that are transferred using GPRS. The beginning of each feeder is also equipped with a smart meter and each feeder is associated with voltage and current measurements at the transformer. So, the total number of voltage measurement points is 81. It is important to note that the phase identification of all smart meters is known, in order to evaluate the performance of the algorithms.

B. Voltage measurement test set

To deal with night-exclusive meters, the phase identification is done with night measurements. The test set is composed of night measurements from all houses and transformer from September 15th, 2017, 23:00 to September 18th, 2017, 4:00, without any measurements from 7:00 to 23:00. This corresponds to 239 time series of 1260 one-minute voltage measurement.

C. Performance measure and empirical assessment

In order to assess the validity of the phase identification, two cases need to be distinguished: (i) an empirical verification (the true phase identification) is available or possible, at a potentially high cost, and can be used to verify the results, (ii) no information is available.

With a (partial) solution, two accuracy metrics can be used benefiting from the true solution. The first one considers each pair (measurement-cluster) as an individual element. The second one considers network elements (house or transformer) as indivisible elements and therefore, an element is either fully correctly identified or not at all. This second metric is not sensitive to the number of measurements per element, while the first one allows to have score for partial identification. In the following, we only focus on the first metric in order to differentiate single-phase and three-phase smart meters. Moreover, if an empirical verification – by sending a technician for example – is possible, but expensive, it can be used to verify the network at some critical points.

Without a (partial) solution, no clues are available to evaluate the quality of the prediction. In this case, some confidence scores may be extracted from the phase identification process in order to identify what may or may not be correctly identified. For example, instead of using all the measurements, one can divide the measurements into several (overlapping) windows of a given length. The consistency of the predictions is assessed by comparing the results of the phase identification for all time windows. To do so, the number of occurrences on which two measurements are in the same cluster are counted to obtain a confidence measure on the association of these two measurements.

VII. Results for the test set: discussions on the selection of the root

Both algorithms need an initialisation with the measurements from a three-phase smart meter. The current section aims at assessing the impact of the root on the performance of the algorithms. The performance index is equal to the ratio between the number of voltage measurements which are correctly identified and the total number of measurements (i.e. 239 for the full data set). When the root is the meter at the distribution transformer, the performance index is 1 for both algorithms. When it is a house, the performance is between 0.67 and 1 for CKM and remains at 1 for the CMT. When half the number of three-phase meters are converted to single-phase meters by randomly selecting one phase, the performance index lies between 0.67 and 1 for both algorithms, while the performance index remains at 1 when the transformer is the root. This highlights the influence of the root on the performance of the algorithms, and hints that the transformer provides a strong root. Indeed, the voltages at the distribution transformer have a particular importance because they influence all the other voltages of the LV network. To further show this behaviour, the algorithms were tested in more demanding settings. The phase identification was run 100 times. For each run, half the meters were randomly removed and half of the remaining were converted to single-phase smart meters by randomly selecting a phase. Furthermore, the 1-minute-average voltages were converted to 15-minute-average voltages by taking the mean of 15 samples at the time. In one case, the root was the transformer and in the other case, the root was randomly selected among the remaining three-phase smart meters. The mean performance index is shown in Figure 1.

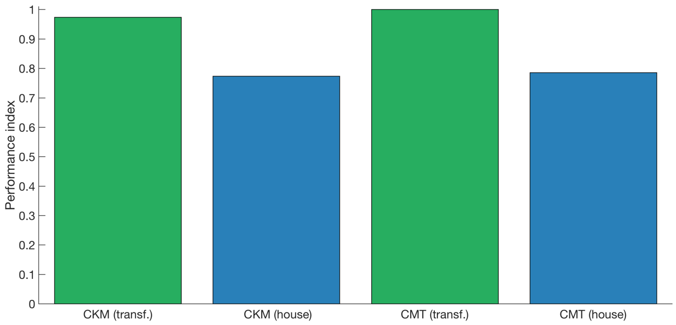


Figure 1. Mean performance indices comparing Constrained k-means (CKM) and Constrained Multi-trees (CMT) in demanding settings (i.e., fewer smart meters among which more are single-phase smart meters, and the sampling period is equal to fifteen minutes).

As can be seen, the mean performance indices are higher with CMT, and for both algorithms the performance is better when the root is the transformer. This is explained thanks to the radial structure of the distribution network and the central role that is played by the distribution transformer. Indeed, the order, with which the measurements are added, and the final tree structures in each cluster are coherent with the network topology, especially with the separation of smart meters between feeders. Since, the CMT algorithm works by sequentially adding the measurement with the lowest cost and making the corresponding tree grow, it may happen that all single-phase measurements are added to the same tree, which grows unequally compared to the two other trees. Indeed, which tree is growing is only selected by the cost and no other method, which may not be optimal. Because of the high number of nodes for that tree, it may occur by chance that the similarity between one of these many nodes is high enough to capture a wrong association and thus gather two phases in the same cluster, leading to three incorrect trees. This could not occur with three-phases measurements because, by definition, one voltage measurement is associated to each tree/phase. This is probably the pitfall of the method and on its own justifies why a central element should be used as root to avoid such a situation as often as possible. Nevertheless, let us note that the multi-trees algorithm cannot reach the optimal solution if the best way to make trees grow is to connect measurements from different phases according to the given distance measure, reflecting the inherent limitation in the data that restrains the performances of the algorithm.

VIII. Performance in different settings

In this section, the performances of both algorithms are analysed when the original data set is modified in four different ways: (i) when one-minute average voltage time series are converted to time series averaging voltages over a longer period of time, (ii) when some random smart meters are removed from the data set, (iii) when three-phase smart meters are converted to single-phase ones by conserving a randomly selected phase, (iv) when the time window over which the phase identification is performed is less than the one from the original data set. Given the results of the previous section, the measurements from the distribution transformer are used as roots of the three clusters. Both algorithms are applied on exactly the same data, and thus results are comparable across methods. In all experiments, the performance index is the one defined in the previous section, i.e. the ratio between the measurements correctly identified and the total number of measurements.

Moreover, all experiments were repeated 100 times in order to obtain average results, since the process is partially randomised, except for the first case where only one experiment is carried out. Finally, all results are represented with box plots: on each box, the central mark is the median, while the edges of the box are the 25th and 75th percentiles. Outliers are identified and plotted individually, while the whiskers extend to the most extreme data points not considered as outliers.

A. Influence of the voltage-averaging period

If the smart meters' internal clocks are well synchronised, averaging the voltages over a short period of times is more reliable because it keeps more variations in the voltages.However, due to technical limitations, some equipment is not able to measure voltages at a high frequency, e.g. instead of one measure per minute, one measure is taken every five minutes, or even every 15 minutes, which is established as the standard in the power system community.

Even if it seems better to gather as much data as possible, it also leads to huge amounts of data that need to be transferred and then stored, potentially inducing a trade-off between the amount of data required to achieve good performances in phase identification, while being reasonable on the amount of data.

Figure 2 illustrates the performance index when the sampling period increases up to one measure per hour. When analysing the performances of the reference method, it appears that the ability to solve the phase identification decreases as the sampling period increases. As expected, the loss of information caused by the lower resolution makes it harder to correctly gather measurements. However, it also shows that (i) the performances of the constrained multi-trees are more stable, and (ii) this method is able to perfectly solve the phase identification problem, even with an extremely high sampling period.

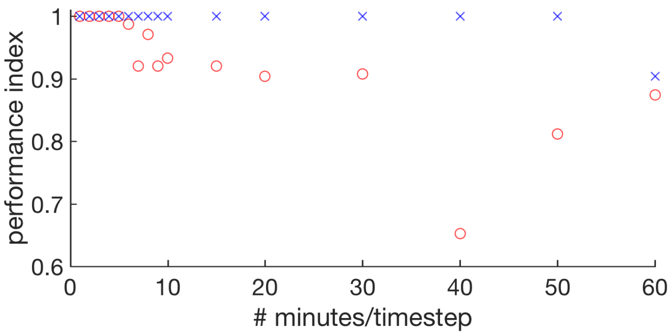


Figure 2. Performance index as a function of the decrease of the sampling frequency. With red circle markers, the reference (constrained k-means clustering) and with blue cross markers, the proposed method (constrained multi-trees) which outperforms the former.

B. Influence of ratio single-phase – three-phase smart meters

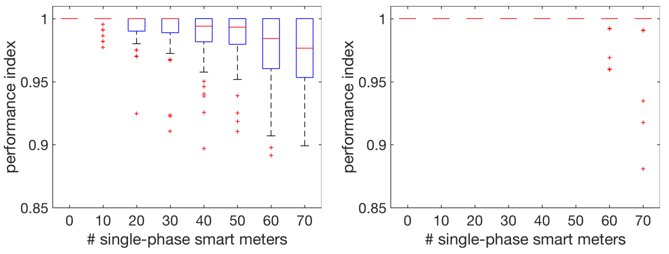
From a correlation point of view, the three-phase constraints may help to find the phase identifications for measurements. Indeed, if, in the three voltage measurements from a three-phase smart meter, two are strongly correlated to the other measurements, while the other one is not. The constraint will ensure that the final measurement is properly identified by correctly identifying the first two. However, if the same voltage measurements were to belong to three different single-phase smart meters, and thus not be constrained, the third measurement could be associated to a wrong cluster because it could be more closely correlated to measurements from the clusters to which the first and second measurements belong. Single-phase smart meters complicate the phase identification process.

Figure 3 displays the performance index when a certain number of randomly selected three-phase meters (the maximum of three-phase meters being 79) are converted to single-phase ones.

From Figure 3, the stability of the constrained multi-trees algorithm seems obvious when the number of single-phase meters increases, while the performance of the reference method decreases significantly. Indeed, measurements from single-phase smart meters are harder to correctly identify, but growing tree structures prove to be effective with a great number of single-phase smart meters.

C. Influence of missing smart meters

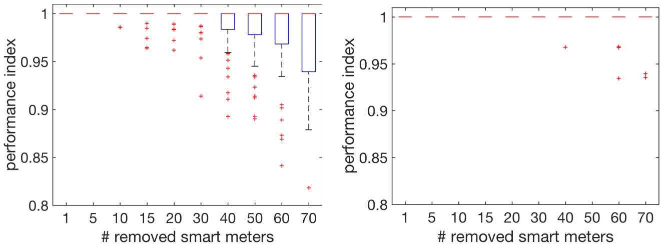
In this section, smart meters (without distinction between single phase and three phase) are randomly removed in order to virtually reduce the observability on the network, and evaluate the critical number of meters that is required to solve the identification problem. This is equivalent to increasing the number of houses that are not monitored.



(a) Constrained k-means (b) Constrained multi-trees

Figure 3. Performance index according to the evolution of the number of single-phase smart meters.

Similarly, as in results from section VIII-B, constrained multi-trees performance are more stable and significantly better than the reference when the data set is small (i.e. less than 40 smart meters). Given that the results for CMT only shows outliers and the median is equal to 1, the outliers probably correspond to cases where the remaining smart meters are not correlated to other measurements, e.g. where a single smart meter is the only one remaining at the end of a long feeder.



(a) Constrained k-means (b) Constrained multi-trees

Figure 4. Performance index according to the number of removed smart meters (i.e., a decrease in the total number of smart meters).

D. Influence of sliding time windows

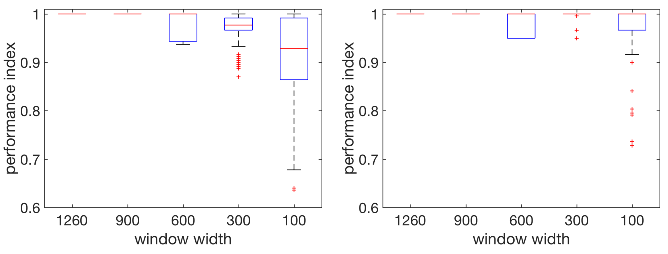
Measurements of the original data set gather 1260 samples each, which may seem enough when measuring similarities thanks to correlation. However, one may ask if the algorithm is able to solve the identification problem with a smaller set of measurements.

To perform this experiment, only a random sub-window of a given length L is kept. In other words, L consecutive measurements were randomly extracted in order to produce a new dataset with a smaller recording time window.

As expected, smaller time windows complicate the task of phase identification. However, constrained multi-trees still manage to achieve better performance with a very small number of consecutive measurements, i.e. a very small measurement time window.

IX. Conclusion

This paper introduces a novel method to identify the phases of smart meters in LV distribution networks by clustering the voltage measurements using graph theory and the correlation between measurements. The algorithm, named constrained multi-tree clustering, successfully manages to identify the phases of smart meters based on real voltage measurements from an LV network in Belgium. It takes, as input, a root smart meter upon which the clustering process is done. Computations show that a good choice for the root is the meter at the distribution transformer. This is mainly due to the central position played by the distribution transformer in LV networks. The performance of the algorithm in various settings is compared to those of constrained k-means clustering. The constrained multi-tree method performs better regardless of the ratio between single-phase and three-phase smart meters, or the increasing number of houses that are not monitored, i.e. a very small set of measurements. Finally, another one of its advantages is its capacity to properly handle voltages that are averaged over a longer period of time without wrongly identifying smart meters.



(a) Constrained k-means (b) Constrained multi-trees

Figure 5. Performance index according to the decrease in the time window width.

In this paper, the algorithms have been tested on voltage measurements from a 3-phase 4-wire network with ungrounded neutral. Future works could include tests on measurements from other network configurations, such as (i) 3-phase 4-wire with grounded neutral, (ii) 3-phase 3-wire (3x230V) and no ground. The proposed algorithm need not be adapted since the identification is based on the fact that voltage measurements between the same phases are more correlated, than measurements between different phases. On the one hand, the performance are expected to increase in case (i) because the neutral voltage is kept low thanks to its repeated connection to the ground. On the other hand, the chance of misidentification should increase in configuration (ii) because the correlation between two phase-to-phase voltage measurements that share a common phase should be higher than the one between two phase-to-neutral voltage measurements.

We could also improve the method to handle which cluster is growing and at what pace, to avoid the growth of a cluster at the expense of the others. Finally, it could be interesting to use this novel method to infer network topology, especially since the tree-structured assumption seems very well adapted to distribution networks.

References

[1] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, “Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data,” in *Proc. of 2016 IEEE Machine Learning and Applications (ICMLA)*, 2016.

[2] A. Dubey, S. Santoso, and A. Maitra, “Understanding photovoltaic hosting capacity of distribution circuits,” in *Proc. of 2015 IEEE Power & Energy Society General Meeting*, 2015.

[3] B. Bletterie, A. Goršek, A. Abart, and M. Heidl, "Understanding the effects of unsymmetrical infeed on the voltage rise for the design of suitable voltage control algorithms with PV," in *Proc. of the 26th European Photovoltaic Solar Energy Conference and Exhibition*, Hamburg, 2011.

[4] S. Cundeva, M. Bollen, and D. Schwanz, "Hosting capacity of the grid for wind generators set by voltage magnitude and distortion levels," in *Proc. of 2016 Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower)*, 2016.

[5] D. Schwanz, F. Moller, S. K. Ronnberg, J. Meyer, and M. H. Bollen, "Stochastic assessment of voltage unbalance due to single-phase-connected solar power," in *Proc. of 2016 International Conference on Harmonics and Quality of Power (ICHQP)*, 2016.

[6] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying Topology of Low Voltage (LV) Distribution Networks Based on Smart Meter Data," *IEEE Transactions on Smart Grid*, 2017.

[7] R. Mitra, R. Kota, S. Bandyopadhyay, V. Arya, B. Sullivan, R. Mueller, H. Storey, and G. Labut, "Voltage Correlations in Smart Meter Data," in *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[8] J. D. Watson, J. Welch, and N. R. Watson, "Use of Smart-meter Data to Determine Distribution System Topology," *IET Journal of Engineering*, 2016.

[9] A. Rajeswaran, N. P. Bhatt, R. Pasumarthy *et al.*, "A Novel Approach for Phase Identification in Smart Grids Using Graph Theory and Principal Component Analysis," in *Proc. of 2016 IEEE American Control Conference (ACC)*, 2016.

[10] H. Pezeshki and P. J. Wolfs, "Consumer Phase Identification in a Three Phase Unbalanced LV Distribution Network," in *Proc. of 2012 Innovative Smart Grid Technologies (ISGT Europe)*, 2012.

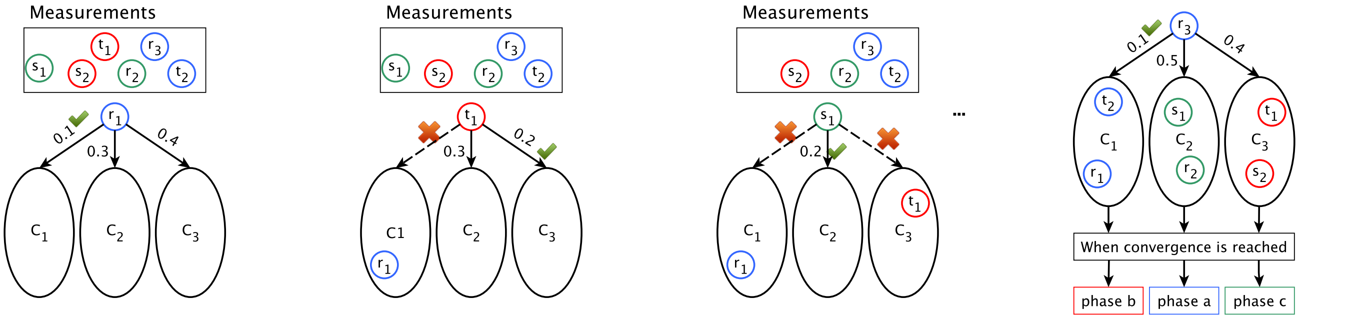
[11] F. Olivier, D. Ernst, and R. Fonteneau, "Automatic Phase Identification of Smart Meter Measurement Data," in *Proc. of the 24th International Conference on Electricity Distribution (CIRED)*, June 2017.

[12] S. Lloyd, "Least Qquares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[13] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means Clustering with Background Knowledge," in *Proc. of International Conference on Machine Learning*, 2001, pp. 577–584.

[14] W. Wang, N. Yu, and Z. Lu, "Advanced metering infrastructure data driven phase identification in smart grid," *GREEN 2017 Forward*, p. 22, 2017.

APPENDIX



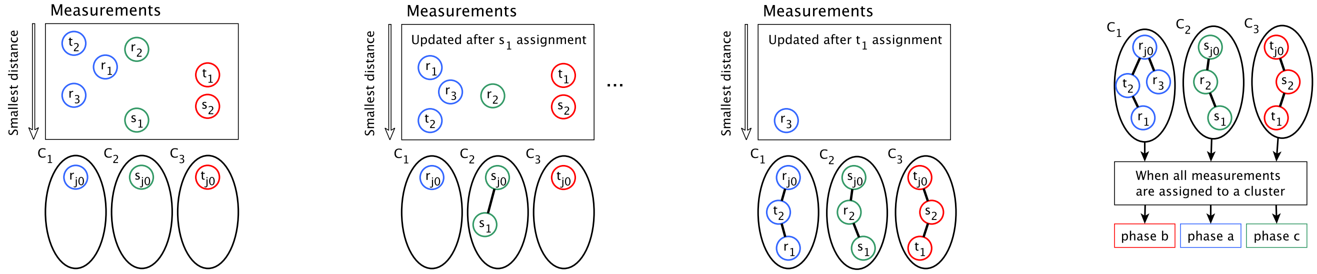
(a) To associate the first measurement from a three-phase smart meter, its distance to all three cluster centres is computed. The closest one (i.e. the one with the smallest distance) is chosen.

(b) The second measurement from a three-phase smart meter is associated based on its distance to the two remaining cluster centres, the other one being unavailable because of the first measurement.

(c) The constraints force the last measurement from a three-phase smart meter to be trivially associated to the remaining cluster.

(d) Since single-phase smart meters do not impose constraints, its measurement is associated to the closest cluster centre.

Figure 6. Schematic representation of the Constrained k-means algorithm applied to phase identification: initially, the cluster centres are defined by the selected root. At the beginning of every iteration, all three clusters are empty. After each iteration, the centres are computed as the average of the measurements associated to the cluster. Measurements are associated to clusters following steps (a) to (c) for three-phase smart meters and step (d) for single-phase smart meters.



(a) The measurements are sorted according to the distance δ_i between them and the closest cluster, while satisfying the constraints and the potential association of the other measurements from the same smart meter.

(b) The measurement with the smallest distance is associated to its potential cluster, expending the tree by being connected to its potential predecessor. The distances δ_i of the remaining measurements are recomputed.

(c) The process of measurement selection and association is repeated until all measurements are in a cluster.

(d) The three clusters are associated to the phases of the network either arbitrarily or by selecting a reference.

Figure 7. Schematic representation of the Constrained Multi-Tree algorithm.