Electrical Power and Energy Systems 144 (2023) 108525

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Phase topology identification in low-voltage distribution networks: A Bayesian approach

Sebastián García [*], Javier M. Mora-Merchán, Diego F. Larios, Enrique Personal, Antonio Parejo, Carlos León

Electronic Technology Department, Escuela Politecnica Superior, University of Seville, Seville 41011, Spain

ARTICLE INFO

Keywords:
Phase identification
Distribution networks
Data analytics
Smart meters
Distribution system operators

ABSTRACT

Knowledge of customer phase connection in low-voltage distribution networks is important for Distribution System Operators (DSOs). This paper presents a novel data-driven phase identification method based on Bayesian inference, which uses load consumption profiles as inputs. This method uses a non-linear function to establish the probability of a customer being connected to a given phase, based on variations in the customer's consumption and those in the phase feeders. Owing to the Bayesian inference, the proposed method can provide up-to-date certainty about the phase connection of each customer. To improve the detection of those customers that are more difficult to identify, after obtaining the up-to-date certainty for all users, the consumption of those who have an up-to-date certainty above a certain percentile compared with the rest of the substation (those that are more likely to be correctly classified) is subtracted from the phase in which they are classified. The performance of the proposed method was evaluated using a real (non-synthetic) low-voltage distribution network. Favourable results (with accuracies higher than 97 %) were obtained in almost all cases, regardless of the percentage of Smart Meter penetration and the size of the substation. A comparison with other state-of-the-art methods showed that the proposed method outperforms (or equals) them. The proposed method does not necessarily require previously labelled data; however, it can handle them even if they contain errors. Having previous information (partial or complete) increases the performance of phase identification, making it possible to correct erroneous previous labelling.

1. Introduction

Knowledge of the current state of distribution networks is essential for correct operation. In this sense, numerous efforts have been made by Distribution System Operators (DSOs) to monitor the grid, especially in secondary distribution networks (low-voltage), which have traditionally suffered major deficiencies in their documentation and monitoring systems compared with transmission or primary distribution [1].

The emergence of Smart Grid technology, such as the Advanced Metering Infrastructure (AMI), is an important milestone in low-voltage networks in the last decade, with the deployment of millions of Smart Meters that allows monitoring of energy consumption profiles, quality of service, and even customer management [2].

However, topological information remains a constraint in secondary distribution networks [2]. This type of information is typically recorded manually in the field by operators at the moment of connection to the grid. A clear example of this limitation can be found in low-voltage European feeder topologies with regard to single-phase customers, in which the actual phase the customers are connected to is usually unknown by the DSOs or, when this information is available, it is not always reliable for reasons such as errors when the connection was registered, unnotified changes, grid reconfigurations after faults or maintenance operations, old networks that have never been documented, etc. [3,4].

Obviously, knowledge of the loads' connection phase in polyphase power networks is important for DSOs to achieve a balanced grid. An unbalanced three-phase system leads to higher technical losses, a reduction in the hosting capacity, problems in service quality (differences among phase voltages that could lead to over/under voltages), reduced lifetime of assets, etc. Moreover, awareness of customers' phase connection is not only important for balancing the grid but it also helps system operators in other situations, such as maintenance operations,

[*] Corresponding author.
E-mail address: sgarcia15@us.es (S. García).

https://doi.org/10.1016/j.ijepes.2022.108525
Received 11 March 2022; Received in revised form 27 June 2022; Accepted 1 August 2022
Available online 26 August 2022

outage identification, non-technical losses detection, etc. [5].

Thus, awareness of the phase topology of low-voltage distribution grids contributes to better operation. In this paper, a novel data-driven phase identification method using load consumption profiles and Bayesian inference is proposed. The proposed method includes a function that uses the energy consumption variations of the customers and phase feeders to obtain the likelihood probability of each hypothesis (phases). With the obtained likelihood probabilities, the current belief in each hypothesis is updated using Bayesian inference. Thus, up-to-date certainty regarding the connection of each customer is provided. To facilitate detection, when a customer is processed, the consumption of customers in the same substation who have a high certainty of being correctly classified is removed from their phase feeders. A percentile-based method was used to obtain a subset of customers that is more likely to be correctly classified. Informative and non-informative prior probabilities can be used, allowing the inclusion of prior knowledge of the phase topology to enhance identification.

The proposed method is intended to be used in 4-wire low-voltage distribution networks in which many single-phase customers (with phase to neutral connection) and three-phase customers (balanced or not) coexist, similar to the European distribution model [6]. This model contrasts with the North American distribution schema, in which most of the distribution is done in medium-voltage, and a single-phase or a delta-connected two-phase transformer is used to supply a single customer or a small number of customers. In the European schema, a delta-wye power transformer with grounded neutral is usually connected to a medium-voltage distribution network and feeds a low-voltage network (usually 400/230 V) in which a large number of customers (usually between 200 and 500) are directly connected. Even though the proposed model is primarily intended to locate single-phase customers in European-like feeders, the model can be easily adapted to other schemes, such as two-phase loads, which are still found in some networks and are very popular in North American distribution networks.

The rest of the paper is organised as follows: Section 2 reviews current state-of-the-art methods for phase identification. Section 3 introduces the problem formulation and methods used in the proposed solution. The phase identification algorithm proposed in this paper is presented in Section 4. The performance and comparison tests are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Literature review

The traditional approach to deal with phase connection identification by DSOs is to perform field-side testing, in which an O&M operator goes to the customer's supply point to check the actual phase in which it is connected. The identification is usually performed by injecting high-frequency signals into each phase line at the secondary distribution substation and verifying which of the signals is received at the customer's supply point [7]. This approach has several disadvantages: it is costly, time-consuming, needs specific hardware and requires well-trained workers. In addition, it is a static method (changes on the grid require to re-evaluate the phase topology) that requires a periodic campaign, which could be a bother for the customers and DSO.

Owing to the drawbacks of using this field-based phase identification method and the increasing popularity of Smart Meters, current efforts to solve the phase identification problems are mainly focused on data-driven methods. Data-driven approaches in the literature can be classified into two main groups depending on the magnitude used: voltage-based and energy-based approaches.

Most voltage-based phase identification methods rely on finding the substation phase feeder that has the highest similarity voltage profile to that of the customer [8]. For example, in [3] and [9], the correlation between customers and feeder voltages was used to identify the network connectivity. Although most solutions seek customer-to-feeder voltage correlation, customer-to-customer voltage correlation can also be used in the phase identification problem [10,11]. Based on the same previous ideas, various authors have used clustering techniques to solve the phase identification problem, as in [12] that used a *k*-means algorithm, or [13] that proposed a cluster algorithm based on a multi-tree structure. Spectral clustering using voltage time series was proposed in [14] and [15]. Clustering algorithms establish the voltages of each phase feeder as representatives of the clusters as initial conditions to increase the convergence. Another phase identification method [16] used active and reactive powers in addition to voltages, and formulated the problem as a maximum marginal likelihood (MMLE). However, it did not yield good results in radial grids (as most distribution networks are, especially in low-voltage European networks). Moreover, the data requirements for this last method cannot be obtained with current metering in low-voltage distribution networks [17].

In summary, voltage-based algorithms generally exhibit good performance [17]. However, they have some drawbacks that make difficult their use over secondary distribution networks. One disadvantage is the need for synchronisation. Voltage measurements must be taken at the same instants of time for correct operation, which is difficult because of clock deviations between devices. This requires dedicated instrumentation with synchronisation techniques, such as Phasor Measurement Units (PMUs). Some methods rely on averaging voltage measurements (e.g., hourly) to dilute timestamp discrepancies. However, the higher the time aggregation, the more difficult it is to find the similarity between customer and phase feeder profile. Unfortunately, these restrictions make the current Smart Meter deployment in Europe [18–20] unsuitable for this approach because Smart Meters are not intended for continuous synchronised voltage sampling nor recording voltage measurements for averaging. These meters commonly store specific events to monitor service quality.

Another approach to solving the phase identification problem is to use active energy measurements, which are broadly available because they are used for billing. In this case, they are fully compatible with the information available through Smart Meters. Energy-based methods can be subdivided into two subgroups based on energy balance and spectral analysis.

The first subgroup is based on a simple concept: the sum of consumption at the secondary distribution substation in one phase must be equal to the sum of consumption of the customers connected in that phase (plus the losses). Relying on the previous idea, [21] established the problem as a regression and used LASSO to solve it (which uses L1 regularization to enforce sparsity in the solution). The same energy-balance concept was used in [22] to solve the phase identification problem with aggregated Smart Meter data. Mixed Integer Linear Programming (MILP) was used in [23] to solve a power flow which minimizes the difference between the estimated and measured power at the feeder head of the secondary distribution substation. In [24] the same authors extended this idea but considered the existence of photovoltaic panels. An approach to identifying low-voltage customers using Kalman filters was proposed in [25]. The authors of this previous paper established three Kalman filters, one for each phase, in which the state variables represent the phase connection of the customer, and the measurement of the filters is the energy delivered by the secondary substation. The use of graph theory and Principal Component Analysis (PCA) has also been proposed for phase identification [4].

However, even though this group of methods use input data that are widely available, they have a serious drawback: their performance decreases as the percentage of unmeasured consumption increases. This unmeasured consumption may be due to different factors, such as technical and non-technical losses (i.e., fraud), measurement errors (e.g., due to instrument tolerance or failure), or simply customers without Smart Meters. Furthermore, three-phase Smart Meters usually report a single aggregated energy measurement; therefore, unbalanced three-phase consumption cannot be disaggregated between each phase.

The second subgroup of the energy-based methods is based on looking for singular variations in customer consumption, which may also be reflected in the aggregated consumption profile of its phase feeder at the head of the secondary distribution substation. This concept was first discussed by Xu et al. [19]. The algorithm used in that study is based on a spectral and saliency analysis, in which representative variations in customer power consumption are identified. Subsequently, the correlation between these variations and those at the phase feeders of the secondary distribution substation are obtained using the Pearson Correlation Coefficient (PCC). The phase with the higher correlation is the candidate phase in which the customer is connected. This approach has two main advantages: it uses energy measurements as input, which are widely available for DSOs, and unlike the algorithms based on energy balance, it can provide good results even if the sum of measured consumptions is not close to 100 % of the delivered power; (because the identification of a customer only depends on its consumption profile and the consumption profile at the feeder head of the secondary distribution substation).

Jimenez et al. [26] improved the algorithm of Xu et al. by introducing a change in the customer's processing sequence and some statistical metrics to evaluate if the number of variations found and the correlation results were significant enough to validate the identification. These changes result in a significant improvement in accuracy with fewer samples and a lower percentage of missing measurements than the initial algorithm. The same authors also published a novel phase identification method based on the same ideas, but using genetic algorithms [27]. This method obtained better results than the previous one, but only when the percentage of measured energy is near 100 %. Hosseini et al. [20] proposed a method based on similar ideas, they performed a high-pass filter on consumption profiles to obtain the high-frequency components to later use them as input of the modified *k*-means clustering algorithm. In the modified *k*-means method, the centroids are predefined as the high-pass filtered profiles of the phase feeders, and the loss function is a linear correlation based on the PCC.

Although this last subgroup of algorithms can obtain good results when having unmetered loads, they always provide a solution without any uncertainty metric to support the results. Furthermore, these algorithms are primarily intended for situations where the entire topology is unknown. Having prior knowledge about the connection of some customers could help in the convergence of the rest if the information of the previously assumed customers is valid. However, if the previous information about the connection contains errors, these would be unrecoverable and would alter the results for all customers.

Bayesian inference has been widely used in power systems [28–30] even in topology-related applications. For example in [31,32], to solve topology errors obtaining the current state of substation switches. However, its use in phase identification has not been researched. To the best of the authors' knowledge, the only Bayesian-based approach used in the phase identification problem was presented in [33]. In this previous paper, the authors used voltage and active and reactive powers to obtain the System State Estimation (SSE) of a bank of possible models; subsequently, they obtained the probability of each model to be correct in a Bayesian way. The approach is simple and has some drawbacks that make its application difficult. It requires the current topology of the network and the location of customers. In addition, the complexity of the problem increases exponentially as the number of customers increases (it needs to evaluate 3^n SSE models in a three-phase system with n number of loads).

In conclusion, some problems have been detected in the current phase identification methods. In this sense, this paper proposes a novel data-driven phase identification method based on Bayesian inference to overcome the disadvantages of the available methods. In summary, the main contributions and advantages of the proposed method are as follows:

- It uses the energy measurements of customers and the feeder head of the secondary distribution substation, which, unlike voltages or other measurements, are collected periodically by the DSO and are widely available.
- The proposed method works even with unmetered loads (low Smart Meter penetration, fraud, etc.) maintaining good performance even when large substations with many customers are processed.
- Unlike other similar previous phase identification methods, and owing to the Bayesian inference, the proposed method can work online. Even though this feature could be considered just an implementation consideration, it enables the possibility of providing up-to-date certainty about the results, which is a feature that no other similar phase identification methods have. Having up-to-date certainty about the results makes possible to cast doubt on them.
- It does not require previously labelled information about the phase topology or any other type of topology-related information. However, if previous information (partial or complete) about the phase labelling is provided, the algorithm can handle it. If previously labelled information is provided, the performance of the algorithm increases substantially even when errors are present in the labels. In addition, erroneous labels are corrected.
- In a worst-case scenario (no previously labelled information), the algorithm outperforms (or equals) other state-of-the-art phase identification methods.

3. Materials and methods

3.1. Problem formulation

Let a three-phase low-voltage network be fed by a secondary distribution substation with N single-phase customers, with M active energy measurements associated with each customer and to each phase feeder at the head of the substation. For this substation, let f_{pm} be the energy consumption at phases p at instant m at the feeder head of the substation with $p \in \{R, S, T\}$, and c_{nm} be the consumption of customer n at instant m. Both f_{pm} and c_{nm} are incremental (i.e., the measure in instant m is the energy consumed between instant m and *m-1*).

For easy management, we create **F** (1) and **C** (2), which are two matrices with the three-phase feeders' consumption per phase and the customers' energy consumption profiles connected at the secondary distribution substation between instant 1 and M.

$$\mathbf{F} = \begin{bmatrix} F_R \\ F_S \\ F_T \end{bmatrix} = \begin{bmatrix} f_{R1} & \cdots & f_{RM} \\ f_{S1} & \cdots & f_{SM} \\ f_{T1} & \cdots & f_{TM} \end{bmatrix} \tag{1}$$

$$\mathbf{C} = \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1M} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NM} \end{bmatrix} \tag{2}$$

Let x_{pn} be a binary variable representing customer n to be connected in phase p. Thus, for each single-phase customer, $\sum_{\mathrm{p}} x_{pn} = 1$ is satisfied. The sparse matrix **X** (3) shows the connection of all loads in the secondary distribution substation.

$$\mathbf{X} = \begin{bmatrix} X_1^t & \cdots & X_N^t \end{bmatrix} = \begin{bmatrix} x_{R1} & \cdots & x_{RN} \\ x_{S1} & \cdots & x_{SN} \\ x_{T1} & \cdots & x_{TN} \end{bmatrix} \tag{3}$$

Using the law of energy conservation, the problem can be represented as (4). Where **E** (3x*M*) represents the unmeasured consumptions. In addition, unbalanced three-phase measurements that are reported as an aggregated single measurement they are included in this term as well because this information cannot be directly associated with any phase.

$$\mathbf{F} = \mathbf{XC} + \mathbf{E} \tag{4}$$

Therefore, the problem is to solve Eq. (4) to obtain **X** and **E** matrices. This can be easily solved as a linear regression problem, using LASSO (for example), which enforces sparsity on **X** owing to L1 penalty regularization [21]. However, as mentioned in the previous section, the performance of these energy-conservation approaches decreases as the unmeasured energy increases (**E** matrix). This major drawback makes these approaches unideal. Nevertheless, as shown in [19,26,27] and validated by this study's results, the correlation between variations in consumer consumption and supply in phase feeders can solve the drawback of the energy-balance approach. Thus, the proposed method does not use the energy conservation law (Eq. (4)) to solve the problem. Instead, the correlation of the consumption variations (frequency domain) will be used in the proposed phase identification method.

The variation of energy consumptions can be calculated as a function of depth k, as the difference in consumption between instant $k+t$ and t for each customer and each phase feeder can be obtained as (5) and (6), respectively.

$$VC_{nk} = \left[c_{n(k+1)} - c_{n1},\ \cdots, c_{n(k+t)} - c_{nt},\ \cdots,\ c_{nM} - c_{n(M-k)}\right] \quad \forall k \in [1, M-1],\ \forall t \in [1, M-k] \tag{5}$$

$$VF_{pk} = \left[f_{p(k+1)} - f_{p1},\ \cdots, f_{p(k+t)} - f_{pt},\ \cdots, f_{pM} - f_{p(M-k)}\right] \quad \forall k \in [1, M-1],\ \forall t \in [1, M-k] \tag{6}$$

Where VC_{nk} represents the variations in the energy consumption of customer n between instant $k+t$ and t. Similarly, VF_{pk} represents the variations in feeder p between instants k and $k+t$. The lengths of vectors (5) and (6) are not fixed and depend on the depth used to obtain the consumption variation (k value). In addition, $M-1$ possibilities of k can be used thus, the same number of vectors ($M-1$) can be obtained. Based on this, it is possible to define the variation matrices of the customers' load profiles and phase feeders, as shown in (7) and (8), respectively.

$$\mathbf{VC} = \begin{bmatrix} VC_{11} & \cdots & VC_{1(M-1)} \\ \vdots & \ddots & \vdots \\ VC_{N1} & \cdots & VC_{N(M-1)} \end{bmatrix} \tag{7}$$

$$\mathbf{VF} = \begin{bmatrix} VF_{R1} & \cdots & VF_{R(M-1)} \\ VF_{S1} & \cdots & VF_{S(M-1)} \\ VF_{T1} & \cdots & VF_{T(M-1)} \end{bmatrix} \tag{8}$$

Where the **VC** matrix represents the variations in the energy consumption of the customers connected to the secondary distribution substation, representing in each row the variation of a customer for all possible values of k (1 to $M-1$). Each column represents the variations of all customers with the same value of k. Similarly, **VF** represents the variations in energy consumption in the three feeders of the secondary distribution substation. Each row represents the variations in a feeder for all possible values of k and each column represents the variations in the three feeders for the same value of k. These matrices represent the salient variations in the energy consumption profiles of customers and feeders. This is similar to the high-frequency part of the load profile signal in the frequency domain. With these variation signals, and considering a customer n connected to a phase p, other related studies [19,20,26,27] have shown that there is a strong correlation between the customer variation profile (VC_n) and the aggregated variation profile at the feeder of the phase in which it is connected (VF_p). A simple example to understand the rationale behind this idea is the next: if a customer has a significant change in their consumption pattern (either reducing or increasing its consumption or generation), that change will be reflected in some way in its phase feeder. As that significant change in the customer will only be reflected in the consumption measurement of its phase feeder, it is possible to locate in which of them the customer is connected by evaluating, for example, the correlation between customer's consumption variations and each of the phase feeders' variations.

3.2. Bayesian inference

One of the main deficiencies identified in the available phase identification methods is the absence of a certainty metric for the result of a customer connected to a certain phase. A possible approach to overcome this is to introduce a way to always obtain an updated probability for each customer to be connected to one phase or another. Bayesian inference can be a powerful tool to achieve this goal.

Bayesian inference provides a method to update the degree of belief associated with a certain hypothesis based on new evidence using Bayes' theorem. Thus, using Bayesian inference, the certainty of a customer connected to a phase can be updated.

Bayes's theorem is shown in (9), where $P(A)$ and $P(B)$ are the independent probabilities for events A and B, and $P(A|B)$ and $P(B|A)$ are the conditional probabilities of A assuming B true, and B assuming A true, respectively.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{9}$$

Bayes' theorem can be rewritten for a better understanding as (10a), where H_i represents hypothesis i and EV represents the new evidence. In Bayesian inference, the term $P(EV|H_i)$ is called likelihood, which represents the compatibility of the new evidence to support the hypothesis. $P(H_i)$ is called *a priori* (or prior) probability, which represents the current probability of H_i to be true before seeing new evidence. $P(EV)$ is the marginal probability, which can be rewritten as the sum of the joint probabilities of EV and all hypotheses (Eq. (10b)). $P(H_i|EV)$ is called the posterior probability, which is the probability of hypothesis i in light of new evidence.

$$P(H_i|EV) = \frac{P(EV|H_i)P(H_i)}{P(EV)} \tag{10a}$$

$$P(H_i|EV) = \frac{P(EV|H_i)P(H_i)}{\sum_i P(EV \cap H_i)} = \frac{P(EV|H_i)P(H_i)}{\sum_i P(EV|H_i)P(H_i)} \tag{10b}$$

Thus, by using Eq. (10b), it is possible to derive the posterior probability of a certain hypothesis as a consequence of the likelihood of new evidence and the prior probability of that hypothesis. In other words, it is possible to update the degree of belief for a given hypothesis under the light of new data.

Coming back to the phase identification problem, the probability of a customer being connected to a certain phase can be updated every time that a new block of data is received by the DSO using Bayesian inference. Therefore, there is no need to use large data intervals to obtain a stable solution but use reduced data windows and update the degree of belief in the solution as the data arrive through the AMI system. This advantage implies that the algorithm could become an online tool which would always consider the most updated version of the available data (evidence).

4. Algorithm description

In this subsection, the proposed phase identification algorithm using Bayesian inference is described. The aim is to estimate the connectivity matrix **X** that relates each customer to a phase. To apply Bayesian inference, it is necessary to define how to construct the prior probability (initial values to start the inference) and the likelihood.

4.1. Prior probability determination

As described above, prior probabilities are the current probabilities of each hypothesis before receiving new evidence. In the first evaluation, when no evidence has yet been considered, and if there is no previous information about any of the hypotheses, the initial probabilities are considered as a uniform distribution (equal probability for each hypothesis). In Bayesian inference, when there is no previous information about the hypotheses, these initial priors are called non-informative *a priori* probabilities.

However, if there is previous knowledge associated with the hypotheses, it is possible to set initial priors based on that knowledge. In this case, it is called an informative prior. This is particularly interesting in the case of the phase identification problem because DSOs may have some previous information in their databases. This possibility could be used to reduce the algorithm convergence time, even if it is not accurate, as evaluated in the next section.

A matrix containing the prior probabilities can be formulated as (11), where each element $P(H_{pn})$ represents the prior probability of the hypothesis of phase p for customer n. If non-informative prior probabilities are considered, then $P(H_{pn}) = 1/3$. Thus, each column of the **PP** matrix contains the probabilities that a customer is connected to each phase, with the sum of the elements in the column equal to one. Similarly, each row represents the probabilities of all the customers of the substation to be connected to one of the phases (e.g., the first row represents the probabilities of customers being connected to phase R).

$$\mathbf{PP} = \begin{bmatrix} P(H_{R1}) & \cdots & P(H_{RN}) \\ P(H_{S1}) & \cdots & P(H_{SN}) \\ P(H_{T1}) & \cdots & P(H_{TN}) \end{bmatrix} \tag{11}$$

4.2. Likelihood determination

Establishing the likelihood function is one of the key steps in Bayesian inference, as it is the way as new evidence modify the current belief on a specific hypothesis. In this study, the new evidence is the energy consumption profile of a customer, specifically its consumption variations, as shown in the previous section Xu et al. [19] proved that the customer's consumption variations show a high correlation factor with its phase feeder consumption variations using the Pearson Correlation Coefficient (PCC). As mentioned in the state-of-the-art revision, this idea of using the PCC was also used by [20,26,27] in their phase identification methods, which showed good results. Thus, a possible way to establish the likelihood of the new data for each hypothesis (three in our case) is to perform the PCC between the variations of a customer and the feeder associated to that hypothesis. As we know, the PCC return values between -1 and 1, where 1 a perfect correlation, 0 no correlation at all and -1 a perfect inverse correlation (as one variable increases, the other decreases). However, probabilities range between 0 and 1. Thus, a relationship between PCC and likelihood probability must be stablished.

As an initial alternative, a linear relationship can be defined ($f(x) = (x+1)/2$) but that would imply that correlation values of 0 (no correlation at all) receive a probability of 0.5, which does not seem reasonable. Another possibility could be a piecewise linear relationship, such as the ReLU function ($f(x) = \max(0,x)$). However, this is not a good approach either because a zero-likelihood probability would be assigned if a negative correlation is obtained. Zero likelihood probabilities would result in a posterior probability value of zero, as can be seen from Eq. (10b), which is not appropriate because it would exclude that possibility forever. Even, it could be possible that the three correlations for each phase would be negative, which would result in an impossible situation after obtaining the posterior probabilities.

As a solution, in order to enhance Bayesian inference, a non-linear function that favors high correlation values was used while penalizing no correlation and inverse correlation results, avoiding giving zero probability results. The equation (12), which is based on the softplus function, is used to establish the likelihood function, where the proposed constants are the best ones that were empirically found.

$$P(VC_n|H_p) = \min\left\{\frac{1}{4}\ln\left(1 + e^{1+4\cdot PCC(VC_n,VF_p)}\right), 1\right\} \tag{12}$$

Where $PCC(VC_n, VF_p)$ is the result of the Pearson Correlation Coefficient between the consumption variations for customer n (VC_n) and variations at the feeder of phase p (VF_p). A representation of Eq. (12) is depicted in Fig. 1. As shown, for small and negative correlation values, a low likelihood is assigned, avoiding zero. Conversely, for medium and high correlations, the relationship is almost linear. Similar to some machine learning models (such as the activation functions of neural networks), this function has the advantage of introducing non-linearities into the problem even though this function does not have a physical interpretation. These non-linearities help enhance Bayesian inference, promoting strong correlations, as demonstrated in the results section.

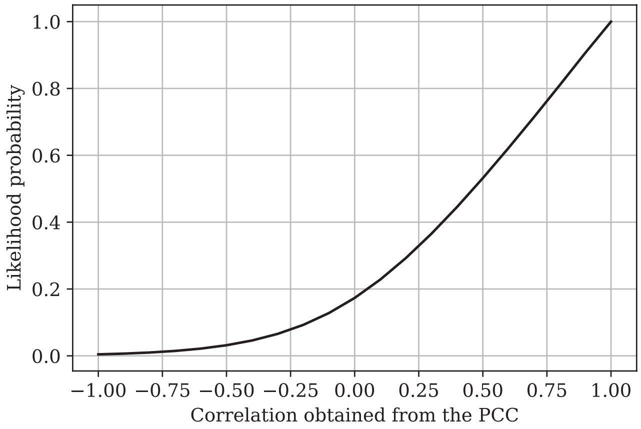

Fig. 1. Representation of Eq. (12) in which the relationship between the PCC and the likelihood probability is obtained.

4.3. Processing sequence

After defining the method for obtaining all the necessary elements to evaluate the Bayesian inference, the method for processing the new evidence (new data) is described in this subsection. In this sense, the data requirements of the proposed method are only the data currently provided by the AMI in European low-voltage distribution networks: the energy measurements from Smart Meters at the customers' point of common coupling and the energy measurements at the phase feeders head of the secondary distribution substation. The formal description of the algorithm is presented in Table 1.

An important feature of the proposed algorithm is its ability to be executed online. For this purpose, the algorithm should be executed periodically each time the AMI receives a meaningful set of data. However, small intervals of execution would result that customers' consumption variations are not sufficiently correlated with their phase feeders. By contrast, large intervals of execution would result in slow certainty update rates. In addition, the granularity of the data interferes with defining the execution interval. Therefore, a trade-off must be established to set the update rate. For example, in Spain, the data are usually retrieved from Smart Meters once a day with measurements at 1 h intervals. In this hourly data schema, it was found that a three-day update rate was a reasonable compromise between obtaining sufficient correlations and a good certainty update rate. Thus, blocks of three days of data were used to update the certainties associated with each hypothesis.

Therefore, once new evidence is acquired to support the hypotheses, and the update time has elapsed, the Bayesian inference is applied. The first time the algorithm is executed (Steps 1 and 2), the likelihood between the variations of each customer and the phase feeders is obtained using Eq. (12). Subsequently, using Eq. (10b), the posterior probabilities are calculated for each customer and hypothesis. The phase with the highest probability is assigned as the candidate solution for each customer (Step 7). Steps 3–6 are omitted in the first iteration of the algorithm. These steps are used to increase convergence; however, they require at least one previous posterior probability.

For the next iterations, after receiving new evidence (Step 3), the obtained posteriors become the new prior probabilities (Step 4). As it was said, after the first iteration, a change in the process is included to increase the convergence. As proven in [19,20,26,27], the correlation of customer variations with its phase increases if other customers' consumption is subtracted from their phases. Therefore, removing some of the customers' consumption in the phase with the highest probability could improve convergence if the hypothesis with the highest

Table 1
Phase identification algorithm.

```
Inputs:
  $I$: vector of length $N$ with the indexes of customers
  VC: variation matrix of customer's energy profiles (Eq. (7))
  VF: variation matrix of feeders' energy profiles (Eq. (8))
  PP: matrix of 3 rows and $N$ columns with the prior probabilities (Eq. (11))
  $\varphi$: percentile of customers with high probability to be considered for subtraction of their consumption
Parameters:
  $J$: vector with the indexes that have a value in its highest probability hypothesis above $\varphi$ percentile.
  X: binary matrix of 3 rows and $N$ columns with the phase connection of all loads in the secondary distribution substation (Eq. (3)).
Algorithm:
Step 1: If the PP matrix has not been passed to the algorithm, initialize it.
  $pp_{pn} = 1/3 \ \forall n \in I, \forall p \in \{R,S,T\}$.
Step 2: If no evidence has been seen before (it is the first time the algorithm is evaluated), wait until evidence are available and then compute the likelihood for each customer and for each hypothesis using (12). Jump to step 7.
  $P(VC_n|H_{pn}) = \min\left\{\frac{1}{4}\ln\left(1+e^{1+4\cdot PCC(VC_n,VF_p)}\right),1\right\} \ \forall n \in I, \forall p \in \{R,S,T\}$
Step 3: Wait until new evidence (VC and VF) are available.
Step 4: Update the prior probabilities in the PP matrix with the posteriors obtained in the last iteration.
  $pp_{pn} = P(H_{pn}) \leftarrow P(H_{pn}|VC_n) \ \forall n \in I, \forall p \in \{R,S,T\}$
Step 5: Extract the value of the hypothesis with the highest probability in the previous iteration for each customer$P(H_n)_{\max}$. Then, extract all indices of customers which its probability is above $\varphi$ percentile and store them in $J$.
  $P(H_n)_{\max} = \max\{P(H_{Rn}),P(H_{Sn}),P(H_{Tn})\} \quad \forall n \in I$
  $J = \left\{i : \left(card(\{i : P(H_i)_{\max} < P(H_n)_{\max}\}) < \phi/100\cdot N\right)\right\} \quad \forall i,n \in I$
Step 6: For each customer in $I$. Obtain a matrix Y with the same shape and structure as X but only considering the phase connection of customers included in $J$. For the rest of customers not included in $J$ and for the customer being processed (n), no connection is assigned to any phase. That matrix is later used to obtain the consumption of customers who are highly probable to be already classified and subtract them from the phase feeders. With the new VF(n) matrix, compute the likelihood for each hypothesis using (12) for customer $n$.
  $\left.\begin{array}{c} \mathbf{Y} = \left[y_{pj} = \begin{cases} x_{pj} & if\, j \in J\backslash n \\ 0 & \text{otherwise} \end{cases}\right]_{3\times N} \\ \mathbf{VF}(n) = \mathbf{VF} - \mathbf{Y}\cdot\mathbf{VC} \\ P(VC_n|H_{pn}) = \min\left\{\frac{1}{4}\ln(1+e^{1+4\cdot PCC(VC_n,VF(n)_p)}),1\right\} \forall p \in \{R,S,T\} \end{array}\right\} \forall n \in I$
Step 7: Compute the posterior probability for each customer of each hypothesis using Eq. (10b) with the likelihood values obtained in the current iteration and the prior probabilities extracted from the PP matrix. Complete X matrix evaluating the phase in which customers are connected as the hypothesis with its higher probability. Jump to step 3.
  $P(H_{pn}|VC_n) = \frac{P(VC_n|H_{pn})P(H_{pn})}{\sum_{\lambda\in\{R,S,T\}} P(VC_n|H_{\lambda n})P(H_{\lambda n})} \ \forall n \in I, \forall p \in \{R,S,T\}$
  $\mathbf{X} = \left[x_{pn} = \begin{cases} 1 & if\ \mathrm{argmax}(\{P(H_{Rn}|VC_n),P(H_{Sn}|VC_n),P(H_{Tn}|VC_n)\}) = H_{pn} \\ 0 & otherwhise \end{cases}\right]_{3xN}$
```

probability is the correct one. To ensure that only customers with correct phase identification are removed, only a subset of customers that are more likely to be correct is used. To select this subset (Step 5), the value of the hypothesis with the highest probability in the previous iteration for each customer is extracted; only those whose probability is above a certain percentile are used, as they are the most likely to be correctly classified. The value of this percentile is a hyperparameter in the algorithm. It was found that the 25th percentile is a good trade-off between selecting customers that are likely to be correctly classified while leaving out the remaining ones. Thus, in the next iterations, for each customer, the consumption of the customers in the subset (excluding itself if it has been included) is removed from their phases with higher probability to obtain the likelihood (Step 6). Subtracting customers with a high probability of being correctly classified from their aggregate-phase feeder consumption causes the algorithm to substantially increase convergence. This is explained by the fact that by removing their consumption variations from the **VF** matrix, it is easier to find the variations of the rest of the customers, which could be more complex to locate because they do not have such significant consumption variations.

5. Use case

In this subsection, the results of the proposed algorithm are evaluated and compared with those of other state-of-the-art phase identification methods.

5.1. Dataset

First, it is necessary to describe the test network used to validate the proposed method, considering that it is primarily intended for European networks (although it would also work in other distribution schemas).

There are numerous test networks that are typically used to evaluate electric power-related algorithms, such as IEEE distribution test feeders [34]. However, these test feeders mostly represent North American distribution networks, which have a different distribution schema than the European ones. As previously described, the North American energy distribution is mostly done in medium voltage, in which many single-phase transformers are used to supply a single customer or a small group of customers [25,35]. This distribution schema contrasts with the European model, in which the secondary distribution (the last mile) is mainly done with a 4-wire low-voltage network (400/220 V) provided by a delta-wye power transformer with a grounded neutral connected to the medium-voltage network [6]. The low-voltage network provided by the power transformer supplies many customers (usually between 200 and 400).

Although there is an IEEE European-like test feeder [34], this network does not adequately represent the European model. The network uses a 3-wire equivalent model using Kron's reduction in Carlson's equation [36], which assumes that multiple groundings are performed along the neutral wire and that the current from the neutral wire returns to the source through the ground connection. This may be a good simplification if balanced or very short networks are considered in which the neutral voltage differences between the secondary distribution substation and customers are low. However, it is not an accurate representation of European networks, since lines can be quite long and follow a TT earth connection in which the ground connection from each customer is independent of the transformer ground and the neutral wire. In addition, this network has only 55 loads, which is quite small compared with the mean size of European low-voltage distribution networks.

Therefore, to evaluate the proposed algorithm, a real European low-voltage distribution network (4-wire with isolated transformer neutral from consumer ground) was used instead of a typical synthetic test feeder. Specifically, the distribution network used was published in [35]. In the paper, the authors describe in detail the model of a real (non-synthetic) distribution network from northern Spain. The grid was composed of 8087 loads distributed among 30 secondary distribution substations. This network has also been used by other authors to test phase identification methods [25].

Table 2
Description of the used substations.

Substation name	ID	N°Customers	% Single Phase
CELLERUELO **(S1)**	C001634	156	85.5 %
EDIFICIO EL MARQUES **(S2)**	C000626	273	87.9 %
PLAZA_ARGUELLES **(S3)**	C000627	369	85.2 %
MARQUESILLA DE CANILLEJAS **(S4)**	C000601	459	89.9 %
VALERIANO LEON **(S5)**	C001007	554	89.7 %
LA GUAXA **(S6)**	C001635	605	88.4 %

For this test, six representative substations of the selected network of different sizes were used to validate the algorithm. A summary of the characteristics of the selected substations is presented in Table 2. This table shows the name of the substation given by the DSO (a symbolic name is also given in parentheses for simplicity), the ID given in [35], the number of customers, and the percentage of them that are single-phase. The six secondary substations have a 22/0.42 kV, 630 kVA delta-wye power transformer with a grounded neutral.

Unfortunately, the consumption profiles provided along with this network only have 20 days of records, which is not sufficient for these types of analyses. Therefore, the consumption profiles used in the model were replaced with real (non-synthetic) hourly consumption profiles from the Medina Garvey database (Spanish DSO). With this information, the model was simulated using OpenDSS.

5.2. Performance of the proposed phase identification algorithm

The proposed phase identification approach was evaluated using three months (90 days) of data. The evolution of accuracy for the six substations in Table 2 is shown in Fig. 2. As can be seen, the algorithm obtains a good final accuracy, with a slower dynamic in substations with a higher number of customers. In addition, the network was also simulated considering some customers without Smart Meters, which represents substations with loads without associated measures. These loads without associated measurements have been simulated considering that just a percentage of customers have smart meters (percentage of Smart Meter penetration term in Fig. 2). Thus, accuracy results are just considering the identification of customers that have a measurable consumption. The effect of these non-measured customers (shown by the

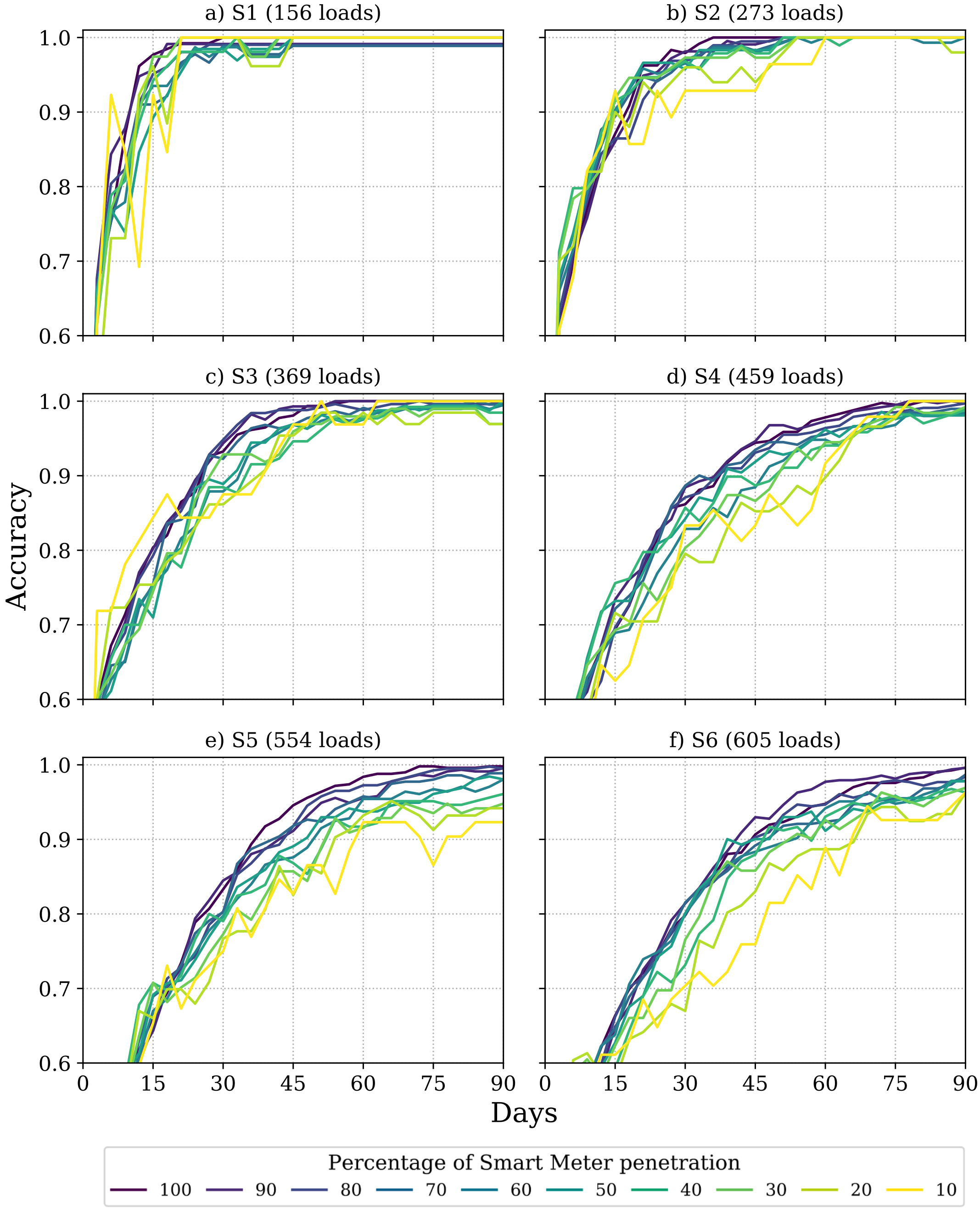

Fig. 2. Accuracy evolution for the phase identification problem using the proposed approach with the customers of substations in Table 1 and with different percentages of Smart Meter penetration.

color in Fig. 2) has little effect on overall performance. Even considering this situation, the proposed phase identification method has good performance, with accuracies higher than 97 % in almost all cases. This is especially important because, although DSOs are increasing the penetration of Smart Meters in their networks, reaching 100 % does not imply that all consumptions are known. This is because there are still unmeasured consumptions (technical and non-technical losses) that cannot be directly used in phase identification. Thus, consider a percentage of customers without Smart Meter is similar to considering an increment in unmeasured consumption.

As previously mentioned, one of the advantages of the proposed algorithm is that it can be run online. In other words, it can update the belief on each possible hypothesis (in which phase a customer is connected) based on new evidence (new data). This is done thanks to Bayesian inference. Thus, this approach can provide an up-to-date certainty of the phase connection of each customer, that is, the proposed algorithm provides the phase connection for each customer, and the confidence of the algorithm in this estimation (for all three phases). This is a clear advantage over other algorithms, making it possible to assess whether the estimation is good. To the best of the authors' knowledge, the previously proposed algorithms of this class do not offer this type of information and do not have any metric on which to rely on for the certainty of the results.

To demonstrate the value of this information, Fig. 3 shows the evolution of the probability for each hypothesis for six customers of S4 (the IDs shown in the titles correspond to the [35] OpenDSS model). These six customers were correctly classified after three months; however, the figures on the right (Fig. 3b, 3d, and 3f) were misclassified on day 30. However, on day 30, the probabilities of the misclassified ones did not show a clear preference from one hypothesis to the others, which made

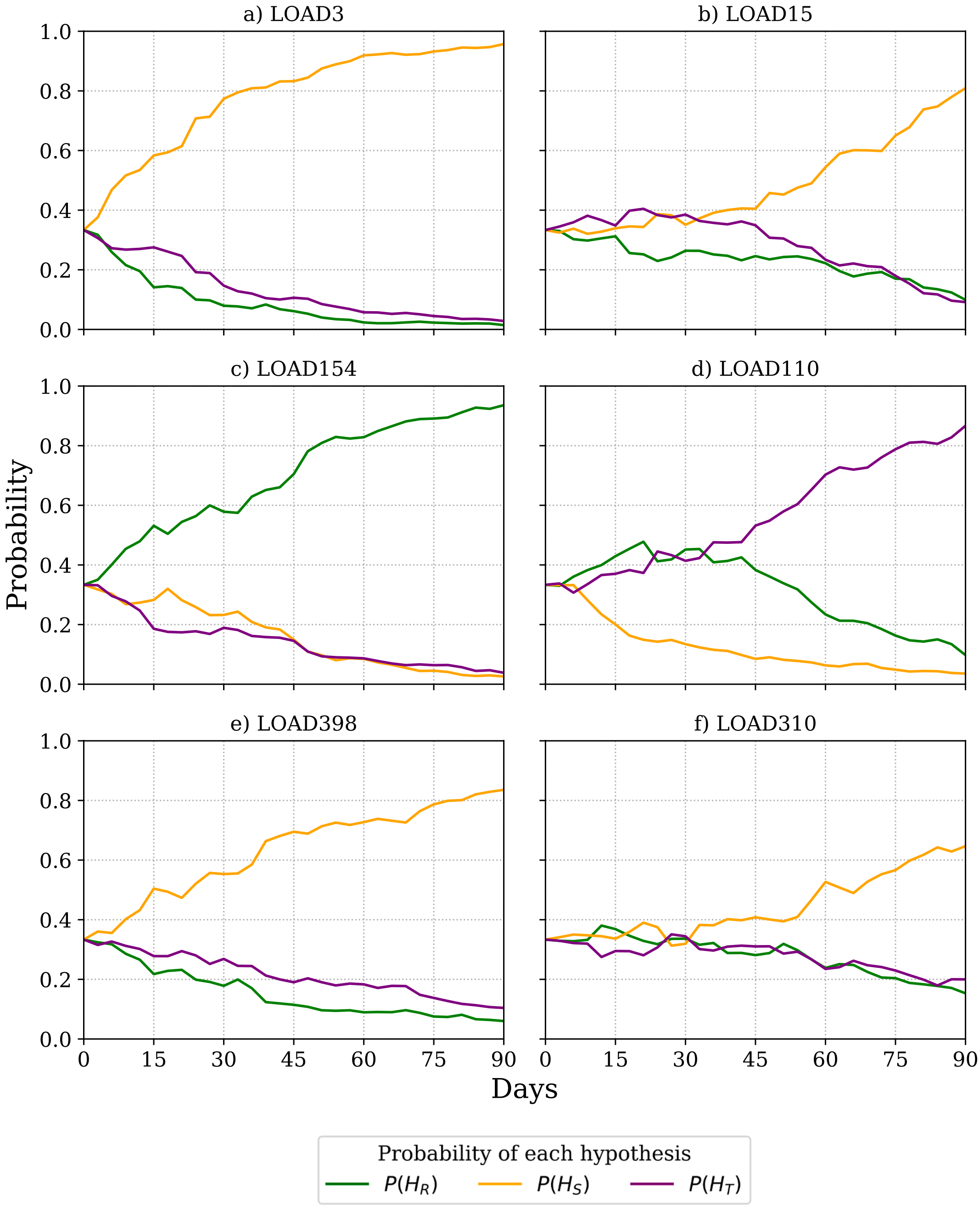

Fig. 3. Evolution of probability for each hypothesis for six different customers. Customers on the left (Fig. 3a, 3c and 3e) are correctly classified from the beginning. Customers on the right (Fig. 3b, 3d and 3f) are misclassified on day 30 but correctly classified on day 90.

them questionable. However, for the correctly classified on day 30 (Fig. 3a, 3c, and 3e), there is a clear preference for one hypothesis. This provided valuable information regarding the certainty of the results. To the best of the authors' knowledge, no other phase identification algorithm provides this type of information.

5.3. Influence of measurement error

To evaluate how the measurement error introduced by the Smart Meters affects the performance of the proposed method, a series of experiments with different levels of measurement error were performed. The current accuracy requirements for Smart Meters deployed in Europe follow IEC 62053-21 [37]. In this sense, the Smart Meter accuracy is not less than Class 2 (2 % error), being usual for manufacturers and DSOs to install Class 1 (1 % error) or lower.

Gaussian noise was added to the dataset to emulate the measurement error. In particular, a Gaussian normal distribution proportional to each measurement was introduced, similar to other studies [4,21,23]. Consideringc_{nm}, the measurement without noise for customer *n* at instant *m*, the erroneous measurement $\widetilde{c}_{nm}$ can be obtained as (13).

$$\widetilde{c}_{nm} \sim N(\mu = c_{nm}, \sigma = \varepsilon c_{nm}/k) \tag{13}$$

Where the mean (μ) is equal to the noiseless measurement, and the standard deviation (σ) is related to the range of the error considered (ε) in the measurement and the sigma rule k. A k value of 2 was considered in the following experiments (95 % of the values within the error considered). Similarly, noise has been introduced into phase feeders (f_{nm}).

To evaluate how the measurement error affects the performance of the proposed method, the accuracy after 30, 60, and 90 days of data were recorded for different levels of measurement error (0 %, 0.2 %, 0.5 %, 1 %, 2 %, and 5 %). The results of the experiments are listed in Table 3.

As shown in Table 3, the proposed method does not show significant changes in its performance under normal levels of measurement error in Smart Meters, with minor changes due to the random nature of the error being added. Just for a measurement error of 5 %, the effect in the performance of the proposed method is clearly visible. However, even with a measurement error of 5 %, the method still exhibited good performance, achieving 98.59 % accuracy after 90 days.

Considering that Smart Meters have error levels less than or equal to 2 %, which is a conservative value, it can be concluded that the proposed method does not have any pitfalls with regard to measurement errors under the tests performed. The method's performance is similar to that obtained in a noiseless scenario.

5.4. Comparison with other published phase identification methods

The accuracy of the proposed algorithm was compared with similar algorithms proposed in the literature. Specifically, [21,19], and [26] were used. These three algorithms are based on energy measurements (as the proposed in this paper), being the first energy balance-based and the rest based on spectral and saliency analyses. Substation S5 and 90 days of data were used for comparison.

Table 3
Accuracy of the proposed method after different periods of time (columns) and under different levels of noise (rows).

Measurement Error	30 days	60 days	90 days
0 %	83.29 %	98.39 %	99.79 %
0.2 %	84.90 %	97.58 %	100 %
0.5 %	82.49 %	96.17 %	99.59 %
1 %	85.71 %	97.58 %	100 %
2 %	82.89 %	97.58 %	99.59 %
5 %	77.06 %	95.37 %	98.59 %

The final accuracy results for different percentages of Smart Meter penetration are shown in Fig. 4. As well as in Section 5.2, the percentage of smart meter penetration represent the customers that have Smart Meters installed in the secondary distribution substation. As can be seen, the proposed phase identification method outperforms [19] and [21], whereas it has a very similar accuracy performance after 90 days with the algorithm proposed by Jimenez et al. [26]. Therefore, a more detailed comparison of this last method and the proposed method was performed.

To further compare the proposed method, the evolution of accuracy was obtained for the six substations listed in Table 2. However, clarifications must be made before making a comparison between the proposed method and the one used as a reference. The available phase identification methods from the literature (including Jimenez et al. [26]) require a block of historical data (there is no way to update the accuracy result and see the evolution upon new data). Therefore, there is no direct comparison with the proposed method. However, it is possible to execute these algorithms multiple times using different sizes of input data to determine how the accuracy evolves and to obtain a set of graphs like the ones at Fig. 2.

For better understanding, instead of representing a multitude of graphs, Fig. 5 shows a comparison of the proposed algorithm with the one that showed the best performance (together with the proposed in this paper) in the previous analysis (Jimenez et al. [26]). For the sake of representation, only 100 %, 90 %, and 80 % of Smart Meter coverage has been plotted. Each point represents the accuracy for the same period elapsed. Three months of data (90 days) in total were used to compare both algorithms. The y and x axes represent the accuracies obtained by the reference and proposed algorithms, respectively. Points below the gray line represent better performance for the algorithm proposed in this paper than the reference algorithm for the same elapsed time (the more distant from the gray line, the more difference between algorithms), whereas points above the gray line represent the contrary.

As can be seen, the proposed algorithm performs better than the reference algorithm on medium and small size substations (Fig. 5a, 5b, 5c, and 5d) where it performs faster up to 90 % accuracies. After that, both algorithms performed similarly. On larger substations, both algorithms perform nearly the same (Fig. 5e) or with a slight difference in favor of the reference algorithm (Fig. 5f). Nevertheless, the proposed algorithm still has the advantage of providing a level of certainty about the result by means of probabilities, with which misclassified customers can be put in doubt as described before.

5.5. Performance of the proposed algorithm considering prior knowledge

The results given until this point were obtained by considering non-

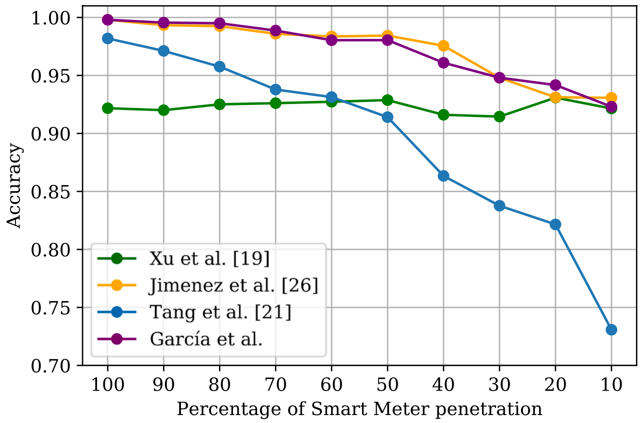

Fig. 4. Accuracy for different percentages of Smart Meter penetration comparison between other state-of-the-art phase identification methods and the proposed in this paper in S5 and using 90 days of data.

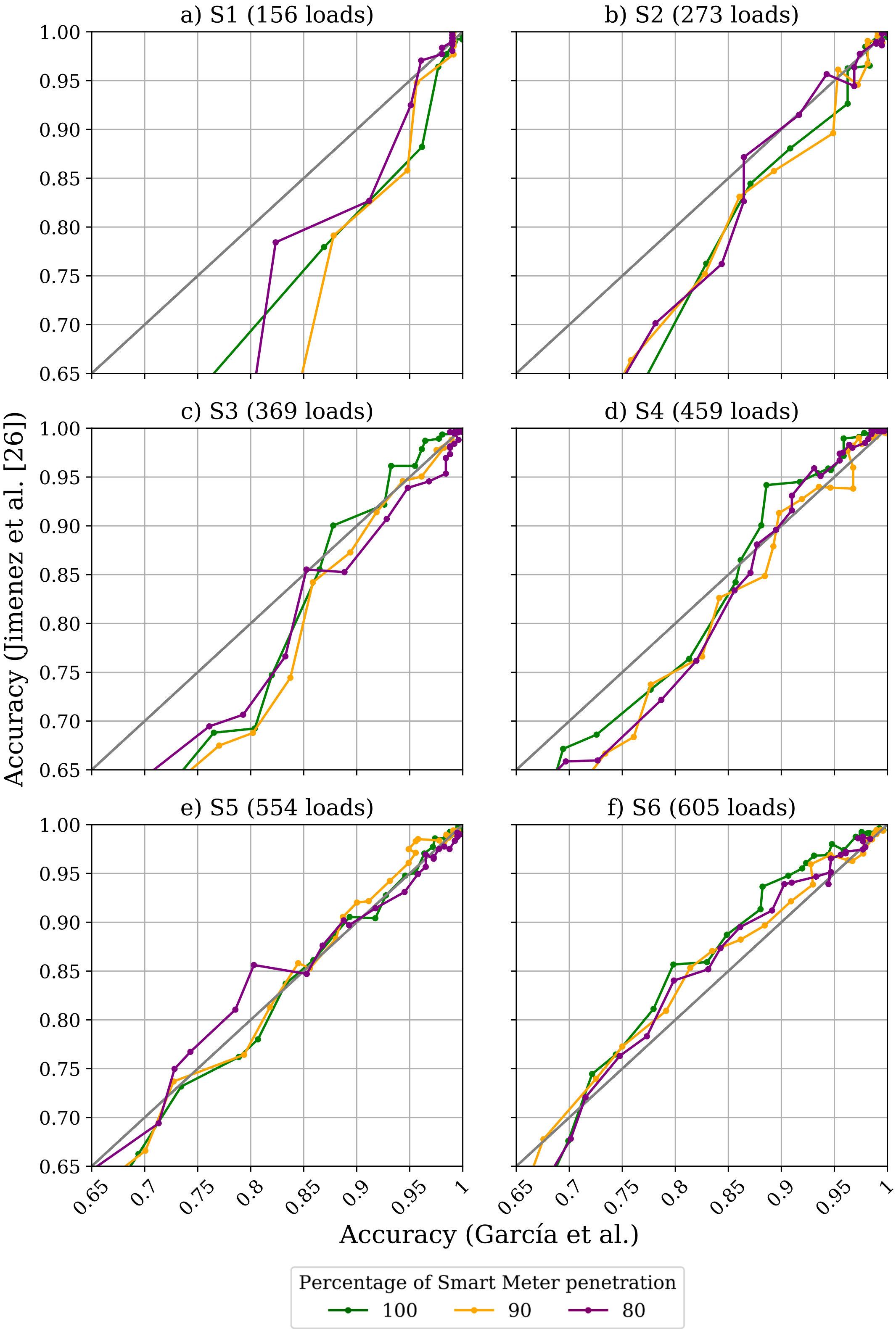

Fig. 5. Accuracy comparison between this paper's proposal and the proposal of Jimenez et al. [26] for the substations of Table 1.

informative initial prior probabilities. Thus, no initial information (true or false) about the phase in which the customers are connected is provided. However, as described in Section 6, if the DSO has any previous information about the customers, it is possible to set initial prior probabilities to consider that information, that is, assigning a higher probability to the phase in which the DSO has the customer registered. However, DSO information is incomplete in many cases and is not always reliable. Thus, to consider all these possible scenarios, the proposed algorithm was evaluated in Substation S4, considering different percentages of customers that have initial information with different degrees of validity of that information and levels of Smart Meter penetration. Previously known customers (being their information valid or not) are assigned a probability of 0.4 to the phase declared in the DSO documentation and 0.3 to the others. Fig. 6 shows the time required to achieve 95 % accuracy for the proposed algorithm. Four different situations were evaluated for the accuracy of the phase connections previously known by the DSO: 0 % (no error), 10 %, 20 %, and 30 % errors. The last column of each graph represents 0 % of the known connections, that is, no information about the phase topology of the substation is previously known. Therefore, this column can be used as a reference.

As can be seen, the performance of the proposed algorithm increases with the presence of previous information, being more noticeable when the information error is low or non-existent (Fig. 6a and 6b). Further, when the presence of error is relatively high (Fig. 6c and 6d), the improvement is noticeable with medium and high percentages of measured customers (40–100 %), being the improvement slightly better in some cases or unnoticeable with low percentages of measured customers (10–40 %). It is important to note that the proposed algorithm performs well even with incorrect previous information. In addition, these tests also show that it can correct the erroneous information given. Even with up to 30 % erroneous prior information (Fig. 6d), the algorithm still achieved 95 % accuracy (so it has corrected the wrong information).

Moreover, the experiment presented in this subsection demonstrates one of the advantages of the algorithm: its ability to operate online. Introducing informative prior probabilities (i.e., certainty values for the hypotheses) is the same as considering that the method has been running for a certain period of time and it has already identified these customers, obtaining the probabilities for each of the hypotheses. In this situation, if a new group of customers is added to the grid, the phase connection of these new customers can be verified within a few days. In contrast, the existing phase identification methods would have to wait to collect more data about these new customers (typically two or more months).

6. Conclusions

Awareness of the customers' phase connection in low-voltage distribution networks gives DSOs useful information to help in the management and operation tasks of their grids. Unfortunately, field-side phase identification methods are costly and time-consuming; therefore, data-driven approaches have become more popular in recent years. As was explained above in the state-of-the-art revision, the available data-driven approaches have some drawbacks that, to the best of the authors' knowledge, have not yet been solved by any previous method in the literature.

In this paper, a novel data-driven phase identification method is proposed. The proposed solution does not require external hardware nor changes in the current distribution networks because it is based on

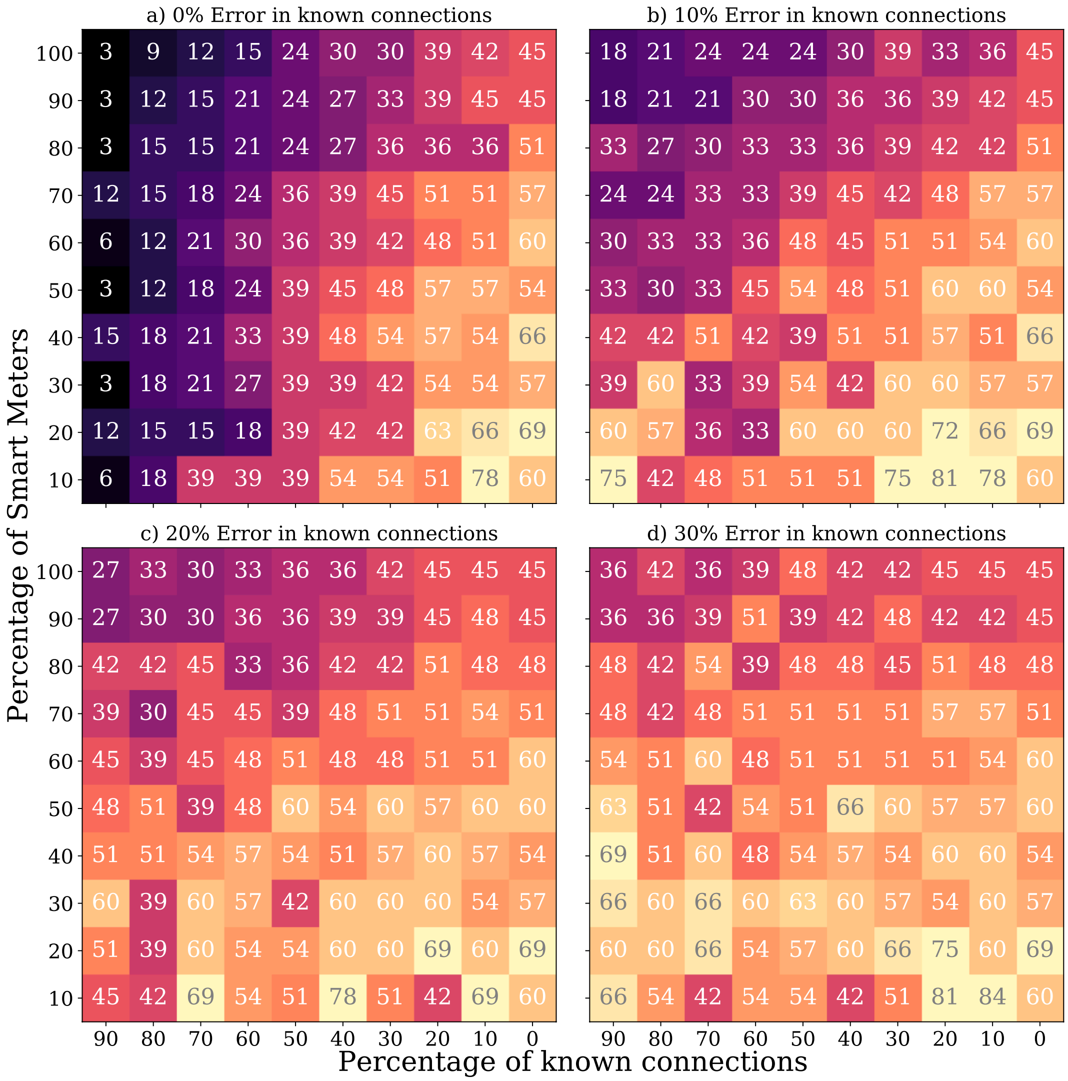

Fig. 6. Time in days to achieve a 95 % of accuracy for the proposed algorithm under different levels of previously known connections, Smart Meter coverage and possible errors in the known connections.

energy measurements (which are already collected by the DSOs). In addition, the algorithm is based on Bayesian inference which, in contrary with current phase identification methods, it allows one to have an up-to-date certainty, providing valuable information about the solution of the phase identification of each customer, making it possible to cast doubt on the misclassified ones. Moreover, in contrast with other similar methods, the method proposed in this paper can handle previous labelling (even if they have errors).

The performance of the proposed phase identification algorithm was tested on six secondary distribution substations of a non-synthetic low-voltage distribution network. The results show good performance under small and large substations and even in situations with low Smart Meter penetrations (unmeasured consumption), with accuracies higher than 97 % in almost all cases. In addition, a comparison with state-of-the-art phase identification methods was performed, showing that the proposed method outperforms most of them using 90 days of input data. A more detailed comparison with one that showed a similar performance in the previous test shows that the proposed solution can achieve better accuracy results using less data.

Moreover, the results also show that having previous information increases the performance of phase identification, and it is even possible to correct erroneous labelling. Thus, even considering that in a worst-case scenario (no previously labelled information), the performance of the proposed method equals other phase identification methods, when the DSO has previous information (even with errors), the proposed method outperforms other similar state-of-the-art phase identification methods.

In summary, this paper proposes a data-driven phase identification method which outperforms (or equal in a worst-case scenario) the current state-of-the-art methods while addressing the drawback of these methods, adding new functions to provide up-to-date certainty metrics and handling previous knowledge about phase topology (even with errors).

In future work, the authors would like to conduct a deeper investigation into the effects of distributed energy resources and local generation on the performance of the phase identification method. Moreover, the authors are currently investigating an extension of the method to not only identify the phase topology, but also to identify on which line the customers are.

CRediT authorship contribution statement

Sebastián García: Conceptualization, Data curation, Formal analysis, Investigation, Software, Writing – original draft. **Javier M. Mora-Merchán:** Conceptualization, Formal analysis, Investigation, Visualization, Writing – original draft. **Diego F. Larios:** Conceptualization, Formal analysis, Investigation, Writing – original draft. **Enrique Personal:** Methodology, Supervision, Resources, Validation, Writing – original draft. **Antonio Parejo:** Investigation, Validation, Writing – review & editing. **Carlos León:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors thank Distribution System Operator “Medina Garvey” for their collaboration and support. This work was supported by the “*Ministerio de Ciencia, Innovación y Universidades*”, Government of Spain under the project “*Bigdata Analitycs e Instrumentación Cyberfísica para Soporte de Operaciones de Distribución en la Smart Grid*”, number RTI2018-094917-B-I00.

References

[1] Lu S, Repo S, Giustina DD, Figuerola F-A-C, Löf A, Pikkarainen M. Real-Time Low Voltage Network Monitoring—ICT Architecture and Field Test Experience. IEEE Trans Smart Grid 2015;6:2002–12. https://doi.org/10.1109/TSG.2014.2371853.

[2] Wang Y, Chen Q, Hong T, Kang C. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Trans Smart Grid 2019;10: 3125–48. https://doi.org/10.1109/TSG.2018.2818167.

[3] Luan W, Peng J, Maras M, Lo J, Harapnuk B. Smart Meter Data Analytics for Distribution Network Connectivity Verification. IEEE Trans Smart Grid 2015;6: 1964–71. https://doi.org/10.1109/TSG.2015.2421304.

[4] Pappu SJ, Bhatt N, Pasumarthy R, Rajeswaran A. Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data. IEEE Trans Smart Grid 2018;9:5113–22. https://doi.org/10.1109/TSG.2017.2680542.

[5] Ma K, Li R, Li F. Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance. IEEE Trans Power Syst 2016;31:2885–91. https://doi.org/10.1109/TPWRS.2015.2481078.

[6] CENELEC - HD 60364 - Low-voltage electrical installations n.d.

[7] Byun HJ, Zheng YP, Choi SJ, Shon SG. New Identification Method for Power Transformer and Phase in Distribution Systems. Appl Mech Mater 2018;878:291–5. https://doi.org/10.4028/www.scientific.net/AMM.878.291.

[8] Short TA. Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling. IEEE Trans Smart Grid 2013;4:651–8. https://doi.org/10.1109/TSG.2012.2219081.

[9] Golub I, Boloev E, Kuzkina Y, Voropai N, Senderov S, Michalevich A, et al. Using smart meters for checking the topology and power flow calculation of a secondary distribution network. E3S Web Conf 2019;139:01059.

[10] Zhou L, Zhang Y, Liu S, Li K, Li C, Yi Y, et al. Consumer phase identification in low-voltage distribution network considering vacant users. Int J Electr Power Energy Syst 2020;121:106079. https://doi.org/10.1016/j.ijepes.2020.106079.

[11] Zhou L, Li Q, Zhang Y, Chen J, Yi Y, Liu S. Consumer phase identification under incomplete data condition with dimensional calibration. Int J Electr Power Energy Syst 2021;129:106851. https://doi.org/10.1016/j.ijepes.2021.106851.

[12] Wang W, Yu N, Foggo B, Davis J, Li J. Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA); 2016. p. 259–65. https://doi.org/10.1109/ICMLA.2016.0050.

[13] Olivier F, Sutera A, Geurts P, Fonteneau R, Ernst D. Phase Identification of Smart Meters by Clustering Voltage Measurements. In: 2018 Power Systems Computation Conference (PSCC); 2018. p. 1–8.

[14] Ma Y, Fan X, Tang R, Duan P, Sun Y, Du J, et al. Phase Identification of Smart Meters by Spectral Clustering. In: 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2); 2018. p. 1–5. https://doi.org/10.1109/EI2.2018.8582318.

[15] Blakely L, Reno MJ. Phase identification using co-association matrix ensemble clustering. IET Smart Grid 2020;3:490–9. https://doi.org/10.1049/iet-stg.2019.0280.

[16] Wang W, Yu N. Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems. IEEE Trans Power Syst 2020;35:3906–17. https://doi.org/10.1109/TPWRS.2020.2977071.

[17] Therrien F, Blakely L, Reno MJ. Assessment of Measurement-Based Phase Identification Methods. IEEE Open Access J Power Energy 2021;8:128–37. https://doi.org/10.1109/OAJPE.2021.3067632.

[18] IEC 62053-21:2020 Electricity metering equipment - Particular requirements - Part 21: Static meters for AC active energy (classes 0,5, 1 and 2) n.d.

[19] Xu M, Li R, Li F. Phase Identification With Incomplete Data. IEEE Trans Smart Grid 2018;9:2777–85. https://doi.org/10.1109/TSG.2016.2619264.

[20] Hosseini ZS, Khodaei A, Paaso A. Machine Learning-Enabled Distribution Network Phase Identification. IEEE Trans Power Syst 2021;36:842–50. https://doi.org/10.1109/TPWRS.2020.3011133.

[21] TANG X, MILANOVIC JV. Phase Identification of LV Distribution Network with Smart Meter Data. 2018 IEEE Power Energy Society General Meeting (PESGM), 2018, p. 1–5. https://doi.org/10.1109/PESGM.2018.8586483.

[22] Brint A, Poursharif G, Black M, Marshall M. Using grouped smart meter data in phase identification. Comput Oper Res 2018;96:213–22. https://doi.org/10.1016/j.cor.2018.02.010.

[23] Akhijahani AH, Hojjatinejad S, Safdarian A. A MILP Model for Phase Identification in LV Distribution Feeders Using Smart Meters Data. Smart Grid Conference (SGC) 2019;2019:1–6. https://doi.org/10.1109/SGC49328.2019.9056591.

[24] Heidari-Akhijahani A, Safdarian A, Aminifar F. Phase Identification of Single-Phase Customers and PV Panels via Smart Meter Data. IEEE Trans Smart Grid 2021;12: 4543–52. https://doi.org/10.1109/TSG.2021.3074663.

[25] González-Cagigal MA, Rosendo-Macías JA, Gómez-Expósito A. Application of nonlinear Kalman filters to the identification of customer phase connection in distribution grids. Int J Electr Power Energy Syst 2021;125:106410. https://doi.org/10.1016/j.ijepes.2020.106410.

[26] Jimenez VA, Will A, Rodriguez S. Phase identification and substation detection using data analysis on limited electricity consumption measurements. Electr Power Syst Res 2020;187:106450. https://doi.org/10.1016/j.epsr.2020.106450.

[27] Jimenez VA, Will A. A new data-driven method based on Niching Genetic Algorithms for phase and substation identification. Electr Power Syst Res 2021; 199:107434. https://doi.org/10.1016/j.epsr.2021.107434.

[28] Clements KA, Costa AS, Agudelo A. Identification of parallel flows in power networks through state estimation and hypothesis testing. Int J Electr Power Energy Syst 2006;28:93–101. https://doi.org/10.1016/j.ijepes.2005.11.015.

[29] Hosseini S, Sarder M. Development of a Bayesian network model for optimal site selection of electric vehicle charging station. Int J Electr Power Energy Syst 2019; 105:110–22. https://doi.org/10.1016/j.ijepes.2018.08.011.

[30] Moradkhani A, Haghifam MR, Mohammadzadeh M. Bayesian estimation of overhead lines failure rate in electrical distribution systems. Int J Electr Power Energy Syst 2014;56:220–7. https://doi.org/10.1016/j.ijepes.2013.11.022.

[31] Lourenco EM, Costa AS, Clements KA. Bayesian-based hypothesis testing for topology error identification in generalized state estimation. IEEE Trans Power Syst 2004;19:1206–15. https://doi.org/10.1109/TPWRS.2003.821442.

[32] Xu Y, Valinejad J, Korkali M, Mili L, Wang Y, Chen X, et al. An Adaptive-Importance-Sampling-Enhanced Bayesian Approach for Topology Estimation in an Unbalanced Power Distribution System. IEEE Trans Power Syst 2022;37(3): 2220–32.

[33] Huang H, Hu Y, Liu S, Xie L. A recursive Bayesian approach to load phase detection in unbalanced distribution system. In: 2017 IEEE Texas Power and Energy Conference (TPEC), 2017, p. 1–4. https://doi.org/10.1109/TPEC.2017.7868280.

[34] Schneider KP, Mather BA, Pal BC, Ten C-W, Shirek GJ, Zhu H, et al. Analytic Considerations and Design Basis for the IEEE Distribution Test Feeders. IEEE Trans Power Syst 2018;33(3):3181–8.

[35] 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. https://doi.org/10.1016/j.ijepes.2019.105712.

[36] Kersting WH, Phillips WH. Distribution feeder line models. IEEE Trans Ind Appl 1995;31:715–20. https://doi.org/10.1109/28.395276.

[37] International Electrotechnical Commission. IEC 62053-21:2021, Electricity metering equipment - Particular requirements - Part 21: Static meters for AC active energy (classes 0,5, 1 and 2). n.d.