Electrical Power and Energy Systems 143 (2022) 108445

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Identification of the phase connectivity in distribution systems through constrained least squares and confidence-based sequential assignment

M.Á. González-Cagigal [a], J.A. Rosendo-Macías [a,*], A. Gómez-Expósito [a,b]

[a] Department of Electrical Engineering, University of Seville, Spain
[b] ENGREEN Laboratory of Engineering for Energy and Environmental Sustainability, Spain

ARTICLE INFO

Keywords:
Constrained least squares
Gaussian distribution
Phase identification
Smart metering
Distribution grid

ABSTRACT

This paper addresses the customer-phase identification problem in three-phase distribution grids including three-phase customers characterized by aggregated energy measurements. The proposed technique first solves a relaxed problem, in which the binary nature of the variables is ignored, which leads to a constrained, least-squares estimation, using as inputs the active and reactive energy readings provided by the smart meters, along with the energy delivered by each phase at the head of the feeder. With the estimated values of the decision variables, and their corresponding variances, a confidence-based selection technique is then applied for the sequential assignment of the customer with the highest joint probability of being connected to one of the three phases but not to the other two. The performance of the proposed procedure is assessed with five different scenarios in terms of accuracy for increasing number of loads and measurement errors. The robustness of the algorithm is additionally tested in the presence of model errors, and its performance is compared to that of existing methods.

1. Introduction

The quality of the service provided to the customers is one of the most important drivers related to the operation and control of distribution networks. In this regard, it is essential to duly characterize the phase which single-phase clients are connected to. In case this information is missing (or inaccurate), feeder unbalance problems may arise, with the associated difficulties, such as excessive voltage variations, which can even violate the grid codes, or a premature deterioration of the grid assets due to the overheating caused by the increased power losses, etc. Moreover, having an accurate topology information benefits the penetration of renewable energy sources, [1,2], in terms of a better per-phase energy balance.

In this regard, despite the efforts undertaken by distribution companies, they frequently lack enough information about the phase connection of their single-phase customers, owing for instance to network reconfiguration after faults, phase switching derived from improper maintenance, or inaccurate tracking of the true load-to-phase connectivity. In these circumstances, a method must be developed to estimate as accurately as possible the actual phase to which a customer is connected in LV feeders, which constitutes the so-called Customer-Phase Identification (CPI) problem.

The relatively recent literature related to the CPI problem is composed of a remarkable number of works, which can be divided into two main groups, [3]: hardware-based and software-based methods. Regarding the first group, the proposed techniques are based on the use of measuring devices designed for this particular purpose, usually deployed at strategic places for a given period of time [4–7]. To reduce the economical impact derived from the deployment of those devices, several software-based methods have arisen. In this category, voltage measurements are considered by some identification methods, in combination with a correlation-based technique [8], a spectral clustering approach [9], or signal processing [10]. Based on voltage observations, [11] presents a procedure to estimate the topology of underground distribution cables. Regarding parameter identification of signal models, new hierarchical (separable) multi-innovation algorithms can be found, in [12] for multi-frequency signals based on the sliding measurement window, and in [13] for signal modeling by using the measurement information. Also on signal modeling, even an optimal adaptive filtering algorithm is proposed in [14] by using the fractional-order derivative technique.

The information provided by the smart meters has also been widely used to address the CPI problem, [15,16]. A method considering Least Absolute Shrinkage and Selection Operator (LASSO) is proposed in [17]. In [18], a novel approach for phase identification using graph theory and Principal Component Analysis (PCA) is tested. The

* Corresponding author.
E-mail address: rosendo@us.es (J.A. Rosendo-Macías).

https://doi.org/10.1016/j.ijepes.2022.108445
Received 14 June 2021; Received in revised form 11 June 2022; Accepted 21 June 2022
Available online 5 July 2022

possible missing information in smart meter data is dealt with in [19] through a correlation analysis. Dynamic State Estimators (DSEs) based on Kalman Filters (KFs) have been recently used in [20], where three KF formulations are compared.

Since the variables to be determined in the CPI problem are not time varying, some of the techniques traditionally used in parameter identification can be applied in this context. These include the least-squares parameter estimation proposed in [21], for multi-input and multi-output systems, the recursive algorithm presented in [22] for signal modeling, and the dynamical window data approach presented in [23] to characterize the frequency response.

In this work, an equality-Constrained Least Squares (CLS) problem is proposed to deal with the CPI problem, assuming that the available information comprises exclusively the hourly active and reactive energy readings provided by individual smart meters, along with the aggregated energy delivered by each phase at the head of the feeder (secondary substation). The complexities arising from the involved variables being binary are circumvented by resorting to a confidence-based sequential selection procedure, which uses the first and second-order statistics of the estimated parameters to determine the single-phase client with the highest probability of being connected to a certain phase but not to the other two. The selected candidate is accordingly assigned, and the same procedure is repeated as long as unassigned single-phase loads remain. The proposed method conservatively assumes that other electrical magnitudes which might be provided by smart meters, such as voltage readings for each load, are not available. This calls for the adoption of a simplified loss model, approximately relating the impact of each load on the total losses. Moreover, such model can be easily adapted for different informative scenarios (e.g. when the reactive energy measurements are not provided, or the network topology is not available). As the information considered in this paper for the CPI problem is similar to that assumed in several previous works, such as [17,18,20], the performance of all those techniques will be compared for different noise scenarios and number of customers.

The main contribution of the proposed methodology, compared to the identification techniques proposed elsewhere, is that single-phase customers are sequentially assigned to one electrical phase on the basis of key statistical information associated with the variables involved in the problem. Through this approach, the explicit enforcement of binary constraints, inherent to the CPI problem, is avoided. Compared to the previous authors' work [20], the main differences lie in the capability of directly handling three-phase clients, for which only a single aggregate energy measurement is assumed (rather than three separate measurements), as well as the use of CLS rather than a nonlinear KF.

The paper is organized as follows: Section 2 presents the proposed technique to approach the CPI problem; Section 3 provides a brief description of the case study used for testing; in Section 4, the results obtained in five different scenarios are presented and discussed, while the proposed assignment procedure is compared in Section 5 with other published works dealing with this problem; the conclusions are presented in Section 6.

2. Proposed phase-identification technique

In this section, the methodology proposed for the phase identification problem is described (see the flowchart of the whole process in Fig. 1).

2.1. Input data

The information gathered and processed throughout the phase identification procedure is as follows:

- Active and reactive energy measurements at the secondary substation during the sampling interval k, for each phase p, EP^{p}_{Sk} and EQ^{p}_{Sk}, with $p = a, b, c$.
- Active and reactive energy readings from the smart meter i during the sampling interval k, EP_{ik} and EQ_{ik}.
- Identification of the connection type for each customer (single-phase or three-phase), N_s being the number of single-phase consumers and N_t the amount of three-phase loads, yielding a total number of consumption readings $N_c = N_s + N_t$, given that only aggregate readings are considered for the three-phase clients. Note the difference in this regard with respect to the previous work [20].
- Topological information, if available, of the electrical distance between each customer and the secondary substation.

Although, in this work, hourly intervals are assumed (according to the current Spanish regulation for smart meters), the proposed procedure can work indistinctly with any scanning rate. In fact, the lower the measurement latency, the more accurate the results obtained.

2.2. Equality-constrained least-squares formulation

The following optimization problem with equality constraints is considered in this paper to address the CPI problem:

$$\begin{aligned} \min_{x} \quad & (b - Ax)^T W (b - Ax) \\ \text{s.t.} \quad & Cx = d \end{aligned} \tag{1}$$

where the objective function is the weighted sum of the squared components of the error vector $\epsilon = b - Ax$, with $\epsilon \sim N(0, R)$ and $W = R^{-1}$. As the elements of ϵ can be assumed to be independent and identically distributed random variables, the covariance matrix will be of the form $R = \sigma^2 \cdot I$, where I is the identity matrix. The elements involved in (1) are described in the sequel:

- The matrix A, related to the consumption of the clients in the distribution grid, is composed in the general case of two submatrices, as follows:

$$A = \begin{bmatrix} A_P \\ \\ A_Q \end{bmatrix}$$

where A_P has the following structure:

$$A_P = \begin{bmatrix} EP & 0 & 0 \\ 0 & EP & 0 \\ 0 & 0 & EP \end{bmatrix}$$

The $H \times N_c$ matrix EP above comprises the active energy readings EP_{ik}, as described in Section 2.1.
The structure of A_Q is similar to that described for A_P,

$$A_Q = \begin{bmatrix} EQ & 0 & 0 \\ 0 & EQ & 0 \\ 0 & 0 & EQ \end{bmatrix}$$

but the entries of EQ are the reactive energy readings.
Therefore, the dimension of A is $6H \times 3N_c$, H being the number of samples (hours with information). In case only active energy measurements are available, the matrix A_Q is not considered, so that $A = A_P$ (dimension $3H \times 3N_c$).
- The vector b is composed of the energy measurements at the secondary substation, including the three phases, as follows:

$$b = \begin{bmatrix} EP^a_S \\ EP^b_S \\ EP^c_S \\ EQ^a_S \\ EQ^b_S \\ EQ^c_S \end{bmatrix}$$

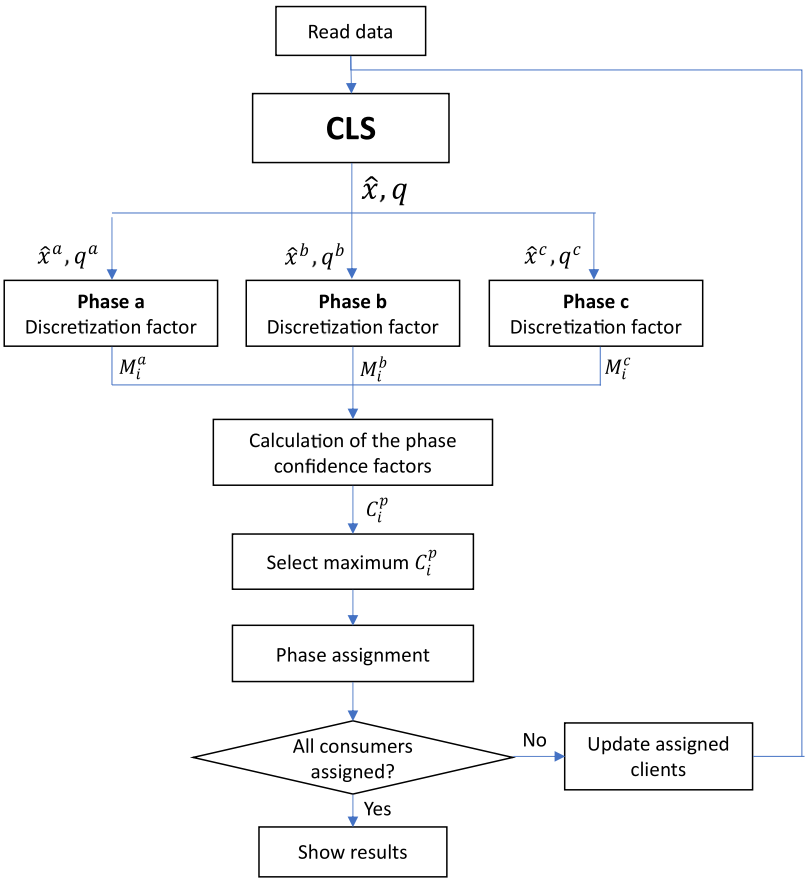

Fig. 1. Flowchart of the proposed methodology.

when active and reactive measurements are considered, or

$$b = \begin{bmatrix} EP_S^a \\ EP_S^b \\ EP_S^c \end{bmatrix}$$

if reactive energy measurements are not provided. It can be noticed that b is a column vector, of size $6H$ (or $3H$).

- The vector x (dimension $3N_c$), to be estimated, also embraces the information of the three phases:

$$x = \begin{bmatrix} x^a \\ x^b \\ x^c \end{bmatrix}$$

The interpretation of each element x_i^p (with $p = a, b$ or c) depends on the type of customer. For single-phase clients, this value should be 1 if the corresponding customer i is connected to the phase p, and 0 otherwise. However, in the proposed method, the binary constraint is initially relaxed so that the conventional CLS technique can be applied, and subsequently enforced through a set of confidence coefficients, as discussed later in Section 2.4. Regarding the three-phase loads, x_i^p represents the fraction of energy consumption associated to phase p throughout the H hourly intervals, with $x_i^p \in [0, 1]$.

- Finally, for both single-phase and three-phase clients, the equality constraint $x_i^a + x_i^b + x_i^c = 1$ should be satisfied, which can be expressed in compact form as $Cx = d$, with

$$C = \begin{bmatrix} I_{N_c} & I_{N_c} & I_{N_c} \end{bmatrix}$$

where I_{N_c} is the identity matrix of size N_c and d is a column vector of size N_c composed of ones.

The above matrix A fully ignores the network losses. In order to enhance the performance of the identification technique, an estimation of the energy losses is added to the corresponding consumption of each client. For the simplified loss model considered, an average hourly current $I_{i,k}$ is calculated for each consumer i over the interval k:

$$I_{i,k} = \frac{\sqrt{EP_{i,k}^2 + EQ_{i,k}^2}}{T \cdot V_n} \tag{2}$$

where V_n is the network rated voltage, and T is the integration or sampling period.

Based on the average current, the active and reactive energy losses attributable to each consumer are accordingly obtained as:

$$EP_{i,k}^{loss} = I_{i,k}^2 \cdot T \cdot r_c \cdot l_i \tag{3}$$

$$EQ_{i,k}^{loss} = I_{i,k}^2 \cdot T \cdot x_c \cdot l_i \tag{4}$$

where r_c and x_c are respectively the conductor resistance and inductance per unit length, and l_i is the estimated electrical distance from load i to the secondary substation.

Those loss terms lead to additional loss matrices A_P^{loss} and A_Q^{loss}, featuring the same structure as that presented previously for A_P and A_Q. Therefore, in the approximate lossy model, the coefficient matrix A becomes,

$$A = \begin{bmatrix} A_P + A_P^{loss} \\ \\ A_Q + A_Q^{loss} \end{bmatrix}$$

2.3. Solution of the constrained least-squares problem

The solution to the optimization problem posed above is provided by the method of Lagrange multipliers. In case all elements of matrix

R are identical to σ^2, as in our problem, the equality constrained formulation reduces to the following system, [24]:

$$\begin{bmatrix} A^T A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \sigma^2 \lambda \end{bmatrix} = G \begin{bmatrix} \hat{x} \\ \sigma^2 \lambda \end{bmatrix} = \begin{bmatrix} A^T b \\ d \end{bmatrix} \tag{5}$$

where λ is the vector of Lagrange multipliers and $\hat{x}$ is the estimate of vector x. Given that G is a nonsingular square matrix, the value of $\hat{x}$ can be obtained from:

$$\begin{bmatrix} \hat{x} \\ \sigma^2 \lambda \end{bmatrix} = G^{-1} \begin{bmatrix} A^T b \\ d \end{bmatrix} = \begin{bmatrix} E1 & E2^T \\ E2 & E3 \end{bmatrix} \begin{bmatrix} A^T b \\ d \end{bmatrix} \tag{6}$$

It can be noticed that, owing to the fact that $R = \sigma^2 \cdot I$, the estimate $\hat{x}$ is not actually affected by the value of σ, unlike its covariance, which is given by:

$$\text{cov}(\hat{x}) = \sigma^2 E_1 \tag{7}$$

2.4. Confidence-based selection and assignment

Under the CLS customary assumptions, each component of vector $\hat{x}$ is considered as a random variable with a normal distribution, $\hat{x}_i \sim N(x_i, q_i)$, where q_i^2 is the respective diagonal of $\text{cov}(\hat{x})$, [24]. In turn, the scalar σ^2, representing the covariance of ϵ, which is required to obtain q_i, can be estimated from the following expression:

$$\hat{\sigma}^2 = \frac{(A\hat{x} - b)^T (A\hat{x} - b)}{f} \tag{8}$$

where f is the number of degrees of freedom, given by:

$$f = 6H - 3N_c + N_c$$

if active and reactive energy readings are provided, or

$$f = 3H - 3N_c + N_c$$

if only active energy measurements are available.

It is worth noting that the real values of the parameters x_i related to three-phase clients might be time-variant, because of the different load distribution of these customers over time for the three phases. However, given that the purpose of the proposed methodology is the identification of the phase connectivity for single-phase loads, only the parameters linked to those consumers will be considered in the sequel for the assignment process.

For the selection method proposed in this paper, the elements of $\hat{x}$ are divided into three N_c-dimension vectors, $\hat{x}^a$, $\hat{x}^b$ and $\hat{x}^c$, with the estimated parameters related to each of the three phases. In the same way, the covariance vector q is split into q^a, q^b and q^c.

Based on the above considerations and assumptions, the normal distribution of $\hat{x}^p$ can be used to quantify the confidence level for a certain single-phase load to be associated with the corresponding electrical phase p and not to the others. Following [20], the cumulative density function over 0.5, denoted as phase discretization factor, M_i^p, provides information about how close a variable $\hat{x}_i^p$ is to the discrete value 1. For this particular case, M_i^p is calculated as follows:

$$\begin{cases} M_i^p = \frac{1}{\sqrt{2 \cdot \pi \cdot q_i^p}} \int_{0.5}^{\infty} e^{-\frac{(x - \hat{x}_i^p)^2}{2 \cdot q_i^p}} dx \\ p = a, b, c \\ i = 1, \ldots, N_s \end{cases} \tag{9}$$

Fig. 2 illustrates the meaning of this coefficient (shaded area in the density function), for a certain parameter $\hat{x}_i^p$.

For each single-phase client i, the information provided by the three coefficients M_i^a, M_i^b and M_i^c is combined to calculate the so-called phase confidence factors, C_i^p (confidence in being connected to phase p and not to the others), using the expressions below:

$$\begin{cases} C_i^a = M_i^a \cdot (1 - M_i^b) \cdot (1 - M_i^c) \\ C_i^b = M_i^b \cdot (1 - M_i^c) \cdot (1 - M_i^a) \\ C_i^c = M_i^c \cdot (1 - M_i^a) \cdot (1 - M_i^b) \\ i = 1, \ldots, N_s \end{cases} \tag{10}$$

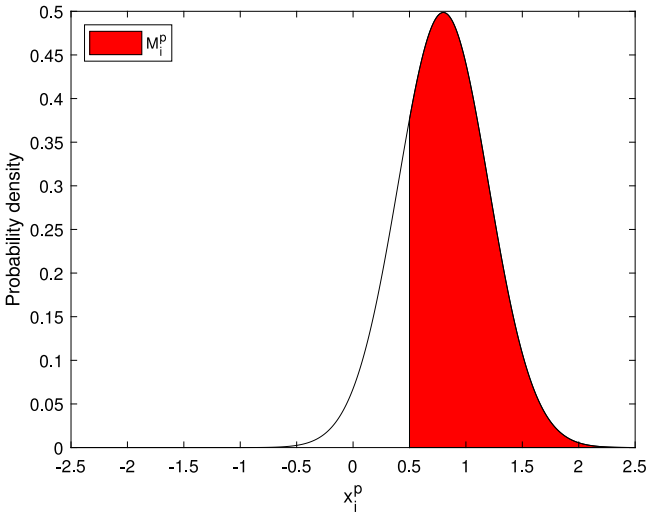

Fig. 2. Graphic representation of M_i^p.

The customer i with the highest likelihood of being connected to the phase p is given by the maximum value of the factors C_i^p. Finally, for the selected client i, the corresponding parameters $\hat{x}_i^a$, $\hat{x}_i^b$, $\hat{x}_i^c$ are assigned integer values (0 or 1).

2.5. Update results

In case all the single-phase clients have been assigned, the process ends and the results are shown. Otherwise, the assigned consumption and its corresponding energy losses are removed from matrix A and introduced as an additional term in vector b, reducing accordingly the number of single-phase loads to be assigned, N_s. Then, the process is repeated with the remaining loads until $N_s = 0$.

3. Case study

For the results that will be presented in Section 4, a set of synthetic distribution grids with a typical European configuration are considered, including N_s single-phase and N_t three-phase customers, which are not necessarily balanced. A particular example of a distribution network is depicted in Fig. 3, with $N_s = 40$ single-phase loads (in blue) and $N_t = 10$ three-phase customers (in red). As illustrated with the small red circles, the information used for the phase assignment is just obtained from the secondary substation and the smart meters installed for all clients in the grid under study.

For our experiments, the active and reactive individual energy consumptions are obtained from [25], where real hourly data from a European distribution company, comprising smart meters readings for 20 days, are provided, leading to a total of 480 energy measurements for each customer. As customers with null consumption provide no information, the corresponding curves are removed from the raw data. For the single-phase customers, the resulting hourly curves are randomly associated to a certain phase (a, b or c), while a random time-varying load distribution is considered for the three-phase clients.

Regarding the resistance per unit length, the value $r_c = 0.223\ \Omega/\text{km}$ has been taken in this work, with a ratio $r_c/x_c = 1$ in all cases. Finally, a load flow is computed at each hour k in order to obtain the energy delivered by each phase of the MV/LV secondary substation. This information fully characterizes the distribution grid model involved in the estimation process.

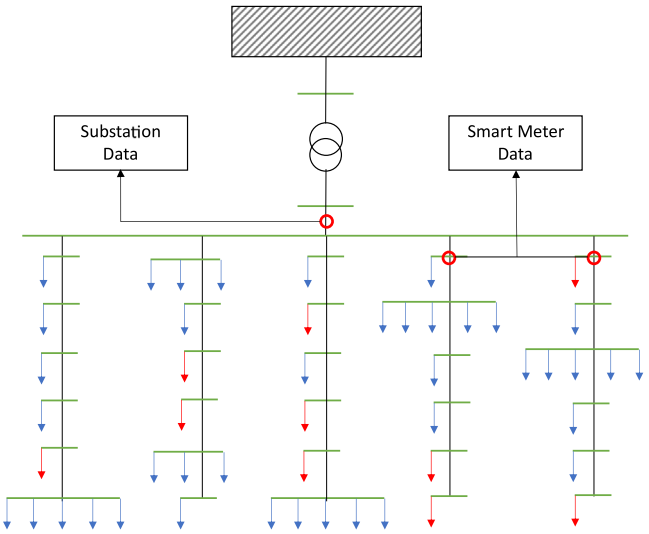

Fig. 3. Single-line diagram of one of the test networks.

Table 1
Estimation results. Scenario I.

Total number of loads	Single-phase clients	Correct assignments	Success rate
50	40	40	100%
100	80	80	100%
200	160	160	100%
300	240	231	96.25%
400	320	292	91.25%
500	400	343	85.75%
600	480	386	80.42%

4. Numerical results

In this section, the proposed sequential assignment technique is tested with several scenarios, which can be grouped into five different categories.

4.1. Scenario I. Original measurements

In the first scenario, the actual active and reactive energy consumptions are considered for the proposed sequential assignment methodology, with no additional errors in the measured values.

The performance of the identification technique is summarized in Table 1 for increasing number of clients in the distribution grid. In all cases, 20% of customers are three-phase.

As expected, the fraction of correct assignments slightly deteriorates as the number of client increases. However, acceptable results have been obtained even for large distribution grids (with 600 loads), with a success rate exceeding 80%.

The good performance of the proposed technique is not only determined by the number of correct assignments, but also by the values of the confidence factors, C_i^p, after each iteration. For a case with 80 single-phase loads (a total of 240 C_i^p factors), Fig. 4 represents the maximum value of those coefficients at each stage of the process, corresponding with the selected single-phase client in each case.

Virtually in all iterations, the maximum value of C_i^p is higher than 0.95, reflecting high confidence in the corresponding assignment.

As an illustrative example, Fig. 5 shows the PDFs obtained for the three estimated parameters, $\hat{x}_i^a$, $\hat{x}_i^b$ and $\hat{x}_i^c$, of a certain single-phase client i. The numeric values of the phase discretization factors, M_i^p, corresponding to the shaded areas, are also provided in the legends. In this particular case, the resulting confidence factors are $C_i^a = 0.024$, $C_i^b = 0.674$ and $C_i^c = 0.003$, implying a higher probability for the customer i to be associated with phase b.

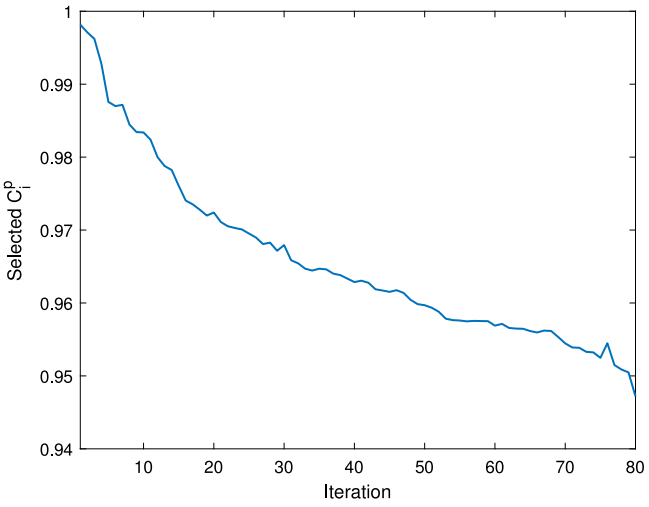

Fig. 4. Evolution of the maximum value of C_i^p.

Table 2
Success rates, in percentage. Scenario II.

Total number of loads	Percentage of three-phase clients				
	10%	20%	30%	40%	50%
50	100	100	100	100	92
100	100	100	100	100	88
150	100	100	100	95.55	85.33
200	100	100	94.28	82.5	74
300	100	96.25	91.90	79.44	68

4.2. Scenario II. Varying number of three-phase clients

The objective of the proposed technique is to provide an accurate estimation of the electrical phase connectivity of the single-phase loads in a distribution network. However, the presence of three-phase loads in the grid, for which only aggregate energy readings are available, might difficult the identification process, given the temporal variability of the load balance in three-phase customers, causing changes in the corresponding regression parameters (note that, in [20], it was assumed that individual phase readings were available for three-phase customers). In this scenario, the performance of the proposed technique is evaluated for increasing shares of three-phase consumers. The success rates for the different cases are summarized in Table 2.

Also as expected, for a given number of loads, the success rate deteriorates with the share of three-phase clients, as their presence is somehow equivalent to increasing "noise" in the available information. Such deterioration might lead to poor results in large grids with high shares of three-phase clients, as in the rightmost bottom case: 68% success rate for 300 loads with an unrealistically share (50%) of three-phase consumers.

4.3. Scenario III. Lack of data

In this scenario, the rate of correct assignments is assessed assuming two less-informative situations, which will be presented separately.

4.3.1. Unavailability of reactive energy readings

In the previous sections, the information used for the identification technique corresponds to 480 hourly readings (20 days), both for the active and reactive energy consumption of each client in the grid. In this case, the performance of the proposed method is assessed when only active energy measurements are available. Accordingly, for the simplified loss model presented in (2)–(4), only the active energy losses

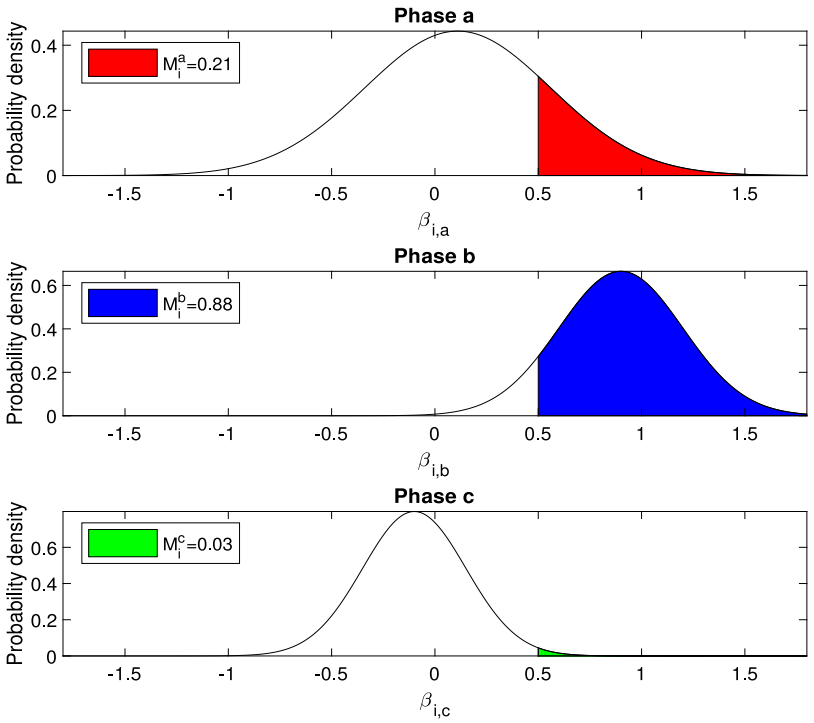

Fig. 5. PDFs for a sample single-phase consumer.

Table 3
Estimation results with only active energy measurements.

Total number of loads	Single-phase clients	Correct assignments	Success rate
50	40	40	100%
100	80	80	100%
200	160	160	100%
300	240	226	94.16%
400	320	279	87.19%
500	400	324	81%
600	480	355	73.96%

are calculated assuming unity power factor ($cos(\varphi) = 1$), yielding the following expression:

$$\begin{cases} I_{i,k} = \frac{EP_{i,k}}{T \cdot U_n} \\ EP_{i,k}^{loss} = I_{i,k}^2 \cdot T \cdot r_c \cdot l_i \end{cases} \tag{11}$$

The remaining assumptions for this case study are similar to those in the first scenario, the corresponding results being summarized in Table 3.

Clearly, the relevance of using both active and reactive energy measurements increases with the number of clients. Nevertheless, the success rates obtained with the proposed technique remain within acceptable values, especially for $N_c < 300$.

4.3.2. Amount of hourly snapshots

The quality of the phase identification is also analyzed when the number of available measurements decreases. For a fixed number of loads, $N_c = 200$, with 20% of three-phase clients, Table 4 presents the estimation results obtained when the number of active and reactive energy snapshots decreases.

The results in Table 4 show the expected deterioration of the success rate for a reduced number of energy readings. The observed impact is less pronounced when both active and reactive energy readings are gathered.

Table 4
Estimation results with decreasing amount of data.

Available data (h)	Success rates with EP and EQ measurements	Success rates with only EP measurements
480	100%	100%
400	100%	100%
300	100%	97.5%
200	100%	90.63%

Table 5
Estimation results for increasing noise levels. Scenario IV.

Noise level	Success rate (%)
2%	100
3%	100
5%	95
7%	91.25
10%	85

A similar study was repeated for different number of clients in the grid. Fig. 6 represents the required number of snapshots to obtain the maximum success rate in each case, with a maximum number of available measurements equal to 480. A roughly linear trend can be noticed for both cases (P&Q or only Q readings), suggesting that more and more snapshots would be needed to assure acceptable results when the number of customers increases, particularly when only P readings are available.

4.4. Scenario IV. Noisy measurements

In this case study, the performance of the proposed technique is tested in the presence of a wide range of measurement errors. In this experiment, a relatively low number of loads is considered: $N_c = 100$, with 20% of three-phase clients. Increasing levels of Gaussian noise are artificially added to the energy measurements after the load flow is computed. Table 5 gathers the success rates for each noise level.

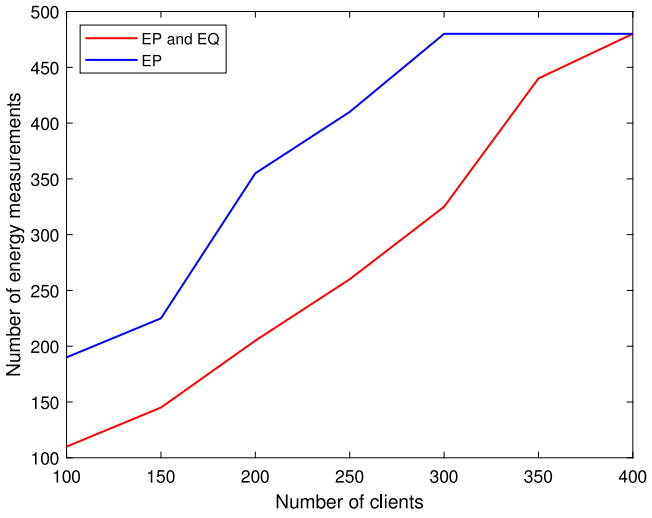

Fig. 6. Required amount of data for the best performance.

Table 6
Estimation results for increasing errors in r_c. Scenario V.

Relative error in r_c	Success rate (%)
5%	100
10%	100
15%	100
20%	97.5
30%	96.25

From Table 5, it can be concluded that the proposed methodology still features acceptable success rates even with measurement errors of up to 10%.

4.5. Scenario V. Model errors

In the simplified loss model considered in the proposed technique, described by Eqs. (2)–(4), there are two main sources of uncertainty: the values of the conductor electrical parameters per unit length, r_c and x_c, and the electrical distance l_i of each client in the distribution grid with respect to the secondary substation.

In this scenario, the robustness of the identification method is evaluated against errors in these parameters. Table 6 summarizes, for the same number of customers as in Scenario IV, the success rates obtained for errors in r ranging from 5 to 30% (assuming $x_c = r_c$).

As can be seen, the proposed technique is not significantly affected by this parameter, yielding success rates over 96% even with 30% errors in r.

When the electrical distance of the customers is not available, an alternative loss model is considered, in which the energy losses associated to each client are proportional to the corresponding consumption [18]:

$$EP_{i,k}^{loss} = \frac{EP_{i,k}}{\sum_{i=1}^{N_c} EP_{i,k}} \cdot \left(\sum_p EP_{S,k}^p - \sum_{i=1}^{N_c} EP_{i,k}\right) \quad (12)$$

$$EQ_{i,k}^{loss} = \frac{EQ_{i,k}}{\sum_{i=1}^{N_c} EQ_{i,k}} \cdot \left(\sum_p EQ_{S,k}^p - \sum_{i=1}^{N_c} EQ_{i,k}\right) \quad (13)$$

where the term $(\sum_p EP_{S,k}^p - \sum_{i=1}^{N_c} EP_{i,k})$ refers to the difference, at instant k, between the energy delivered by the three phases of the secondary substation $(\sum_p EP_{S,k}^p)$, and the total consumption from all customers $(\sum_{i=1}^{N_c} EP_{i,k})$, and the same meaning for the counterpart terms in (13). In case reactive energy readings are not available, only Eq. (12) is used. Table 7 shows the results obtained for increasing number of clients, using the simplified loss model provided by (12)–(13). In all cases, 20% of three-phase loads is considered.

Table 7
Results with no information on the electrical distances.

Total number of loads	Single-phase clients	Correct assignments	Success rate
50	40	40	100%
100	80	80	100%
200	160	160	100%
300	240	224	93.33%
400	320	278	86.87%
500	400	321	80.25%
600	480	350	72.92%

Table 8
Success rates (%) for the different methods considered.

Total number of loads	PCA	LASSO	EnKF	Proposed technique
50	100	100	100	100
100	100	100	100	100
200	85	92.5	95	100
300	79.17	80.83	82.5	93.33
400	67.18	70	65.62	86.87
500	50.25	60.50	53.25	80.25
600	47.29	52.92	48.54	72.92

Compared to the more accurate loss model adopted for the first scenario (Table 1), the performance deterioration is not significant. In all cases, the success rates remain over 85%, giving evidence of the robustness of the algorithm when the exact network electrical model is not available.

5. Comparison with existing methods

Finally, a comparison is made in this section among the proposed technique and three different methods, all of them making use of similar input data for the phase-assignment process:

- A LASSO-based technique [17].
- The method presented in [18], where PCA is considered with exclusively energy consumption curves from smart meters.
- The Ensemble Kalman Filter (EnKF) formulation proposed in [20], which outperformed other KF schemes.

Different scenarios are used for the comparison, with increasing number of clients. In all cases, 2% measurement errors are assumed, with no information available about the electrical distances, which means that Eqs. (12) and (13) are used to estimate the energy losses.

One every five customers is assumed to be three-phase, each with a single energy reading. Given that all the techniques included in the comparison are formulated exclusively with per-phase measurements, the overall consumption for three-phase customers has been equally divided into the three phases. The rates of correct assignments are summarized in Table 8 for all the techniques.

For a reduced number of clients ($Nc \leq 100$), all of the compared methods achieve 100% success rate. As the grid size increases, the PCA, the EnKF and the LASSO-based methods suffer a sharper deterioration of their performance, when compared to that of the proposed technique, which exhibits over 16% better hit rate than the best competing method for $N_c \geq 400$. Those results can be in part explained by the weaker treatment of the three-phase clients with aggregate readings, which is the main advantage of the proposed CLS-based technique.

6. Conclusions

In this paper, a methodology is proposed for the sequential phase assignment of single-phase loads in distribution grids, using smart meter information exclusively. The technique is based on a CLS model and the corresponding statistical distributions of the estimated parameters. At each iteration of the procedure, the load with the highest confidence in being connected to a certain phase is selected, so that the binary equality restrictions are avoided as well as the associated computational problems.

Five scenarios have been presented to assess the performance of the proposed technique, in terms of accuracy and robustness against model errors. The following conclusions can be highlighted from the obtained results:

- The number of correct assignments decreases for increasing number of clients. Nevertheless, a success rate over 80% is still obtained for distribution networks with 600 clients, of which 20% are three-phase.
- For larger shares of three-phase consumers, the performance of the identification process deteriorates, yielding poor results (around 70% hit rate) for feeders with 300 clients and 50% of three-phase loads.
- The sensitivity of the identification technique to the number of customer is higher, as expected, if only active energy measurements are available. The more loads in the same feeder, the more measurement snapshots are needed to perform an acceptable estimation of the phase connectivity.
- Even for measurement errors of up to 10%, the success rates remain over 85% for distribution grids with 200 clients.
- Finally, the robustness of the method has been tested against errors in the conductor resistance and reactance used for the calculation of the energy losses. Additionally, a simplified loss model is considered, with very good results, for the cases where the electrical distances are not available.

The proposed technique has been compared to other published methods, based on PCA, LASSO and EnKF. While the results obtained for a reduced number of clients are similar for all tested techniques, as the network size increases, the proposed technique has shown remarkably higher success rates than the others. This is mainly a consequence of the proposed CLS-based method being the only one to correctly deal with three-phase clients, when only aggregate energy readings are provided. Further research efforts will be devoted to redesigning the proposed procedure, so that additional electrical quantities can be considered, such as voltage magnitudes, which might be available in future distribution systems.

CRediT authorship contribution statement

M.Á. González-Cagigal: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing. **J.A. Rosendo-Macías:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **A. Gómez-Expósito:** Conceptualization, Methodology, Supervision, Writing – original draft, 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.

Acknowledgments

The authors thank the project Solar to Vehicle (S2V, reference INV-3-2021-I-038) and the research project HySGrid+ (CER-20191019) for the financial support.

References

[1] Lueken C, Carvalho PM, Apt J. Distribution grid reconfiguration reduces power losses and helps integrate renewables. Energy Policy 2012;48:260–73.
[2] Melo F, Cándido C, Fortunato C, Silva N, Campos F, Reis P. Distribution automation on LV and MV using distributed intelligence. In: Proc. 22nd int. conf. exhib. electricity distrib. Stockholm, Sweden; 2013.
[3] Hosseini ZS, Khodaei A, Paaso A. Machine learning-enabled distribution network phase identification. IEEE Trans Power Syst 2021;36(2):842–50. http://dx.doi.org/10.1109/TPWRS.2020.3011133, 2020.
[4] Wen MH, Arghandeh R, von Meier A, Poolla K, Li VO. Phase identification in distribution networks with micro-synchrophasors. In: Proc. IEEE power energy soc. general meeting. CO, USA: Denver; 2015, p. 1–5.
[5] Marrón L, Osorio X, Llano A, Arzuaga A, Sendin A. Low voltage feeder identification for smart grids with standard narrowband PLC smart meters. In: Proc. IEEE int. symp. power line commun. appl. South Africa: Johannesburg; 2013, p. 120–5.
[6] Caird KJ. General electric co, meter phase identification. 2010, U.S. Pat. 8, 143, 879.
[7] Kolwalkar AR, Tomlinson HW, Sen B, Hershey JE, Koste GP. Power meter phase identification. 2011, U.S. Patent Application 12/782, 530.
[8] Pezeshki H, Wolfs P. Correlation based method for phase identification in a three phase LV distribution network. In: 2012 22nd australasian universities power engineering conference. Bali; 2012, p. 1–7, 2012.
[9] Blakely L, Reno MJ, Feng W. Spectral clustering for customer phase identification using AMI voltage timeseries. In: 2019 IEEE power and energy conference at illinois, Vol. 2019. Champaign, IL, USA; 2019, p. 1–7.
[10] Pezeshki H, Wolfs PJ. Consumer phase identification in a three phase unbalanced LV distribution network. In: 2012 3rd IEEE PES innovative smart grid technologies europe. Berlin; 2012, p. 1–7, 2012.
[11] Chen CS, Ku TT, Lin CH. Design of phase identification system to support three-phase loading balance of distribution feeders. IEEE Trans Ind Appl 2012;48(1):191–8.
[12] Xu L. Separable multi-innovation Newton iterative modeling algorithm for multi-frequency signals based on the sliding measurement window. Circu Syst Signal Proc 2022;41(2):805–30.
[13] Xu L, Ding F, Zhu Q. Separable synchronous multi-innovation gradient-based iterative signal modeling from on-line measurements. IEEE Trans Instrum Meas 2022;71(1–13):6501313.
[14] Zhang X, Ding F. Optimal adaptive filtering algorithm by using the fractional-order derivative. IEEE Signal Proc Lett 2022;29:399–403.
[15] Arya V, Seetharam D, Kalyanaraman S, Dontas K, Pavloski C, Hoy S, et al. Phase identification in smart grids. In: Smart grid communications. 2011 IEEE International Conference on, 2011, p. 25–30.
[16] Adolfo Gastalver-Rubio, Carmona-Pardo R. Application of neural networks to determine the customer connectivity based on smart meters. Renew Energy Power Qual J 2020;18:8–12. http://dx.doi.org/10.24084/repqj18.ps1.
[17] Tang X, Milanovic JV. Phase identification of LV distribution network with smart meter data. In: 2018 IEEE power & energy society general meeting. Portland, OR; 2018, p. 1–5, 2018.
[18] Satya Jayadev P, Rajeswaran A, Bhatt NP, Pasumarthy R. A novel approach for phase identification in smart grids using graph theory and principal component analysis. In: 2016 american control conference. Boston, MA; 2016, p. 5026–31.
[19] Xu M, Li R, Li F. Phase identification with incomplete data. IEEE Trans Smart Grid 2018;9(4):2777–85.
[20] 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 125.
[21] Ding F, Liu Y, Bao B. Gradient-based and least-squares-based iterative estimation algorithms for multi-input multi-output systems. Proc Inst Mech Eng Part I: J Syst Control Eng 2012;226(1):43–55.
[22] Xu L, Song G. A recursive parameter estimation algorithm for modeling signals with multi-frequencies. Circu Syst Signal Proc 2020;39:4198–224.
[23] Xu L, Xiong WL, Alsaedi A, Hayat T. Hierarchical parameter estimation for the frequency response based on the dynamical window data. Int J Control Autom Syst 2018;16(4):1756–64.
[24] Abur A, Antonio Gómez-Expósito. Power system state estimation: theory and implementation. Marcel Dekker; 2004.
[25] Arpan Koirala, Pablo Arboleya, Bassam Mohamed, Suárez-Ramón, Lucía. Non-synthetic European low voltage test system. Int J Electr Power Energy Syst 2019;118(10):1016.