Application of nonlinear Kalman filters to the identification of customer phase connection in distribution grids

M. A. González-Cagigal, J. A. Rosendo-Macías and A. Gómez-Expósito

Department of Electrical Engineering, University of Seville

Abstract

This paper presents a state estimation approach to address the problem of identifying the phase to which single-phase customers are connected in three-phase distribution grids. The proposed method performs Kalman filtering on the information provided simultaneously by the smart meter of every customer and the aggregated energy consumption measured at each phase of the secondary substation feeding the set of customers. Different nonlinear formulations of the Kalman filter are tested and their performance compared, showing that the ensemble Kalman filter provides better estimation results when the system size increases. The accuracy, robustness and limitations of the estimator are also tested when measurement errors are considered.

Keywords: Ensemble Kalman filter, Cubature Kalman Filter, Unscented Kalman Filter, phase identification, state estimation.

1. Introduction

Correct operation monitoring and control in distribution systems are essential in order to assure a good quality of the service provided to the customers. In this context, it is most important for grid operators to unambiguously know which loads are connected to each of the three phases of the system (the European feeder topology is assumed in this work). An accurate connectivity information is a prerequisite to promote the correct phase balance of LV consumers, alleviating in this way the problems derived from feeder unbalances, such as sharper voltage drops, which can even violate the grid codes, and increasing power losses which also affect the lifespan of the equipment due to temperature rise. Moreover, the penetration of renewable energy resources at the distribution level also benefits from the phase identification, since it helps to establish a better production-consumption balance for each phase of the grid.

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 recording 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 is known as the customer-phase identification (CPI) problem.

The CPI problem has been approached in several ways by previous works. In [1], a signal processing perspective is applied to voltage observations, which are also used both in [2], for a correlation-based methodology, and in [3], where a spectral clustering technique is proposed. The connecting phase of underground distribution transformers is determined in [4] through phase voltage measurements. Smart meters have improved the communication between the loads and the substations and can be also used for the CPI problem, [5], [6]. A method based on Least Absolute Shrinkage and Selection Operator (LASSO) is proposed in [7], also using smart meter data from a LV distribution network. In [8], a novel approach for phase identification using graph theory and principal component analysis (PCA) is tested. The possible missing information in smart meter data is dealt with in [9] through a correlation analysis.

In this work, the CPI is addressed by applying a Kalman filtering (KF) state estimation technique to a set of hourly consumption curves obtained from real loads, along with hourly energy measurements taken at each phase of the secondary substation. The proposed method conservatively assumes that other electrical magnitudes potentially provided by smart meters, such as voltage readings or reactive power consumed by each load, are not available. Moreover, a simplified loss model is adopted, allowing the impact of each load on the energy delivered by each phase of the transformer to be estimated for a given topology.

The KF is a dynamic state estimator (DSE) widely used in electric power systems in any of its diverse forms, [10]-[12]. Particularly, the application of KF to parameter estimation in power systems has been successfully tested, e.g. in [13]-[15], providing evidence of the DSE potential for the CPI problem.

In this work, three nonlinear KF schemes are tested

and compared, namely the so-called unscented Kalman filter (UKF), the cubature Kalman filter (CKF), and the ensemble Kalman filter (EnKF). The proposed technique takes advantage of the hourly information provided by smart meters to sequentially assign customers to the most likely phase. The proposed KF implementation does not explicitly enforce, beforehand, binary constraints for the state variables, but rather adopts a novel, statistically-based inference logic successively rounding state variables to their nearest binary value. As the input information considered in this work is similar to that assumed in [7] and [8], the KF-based estimation is also compared to those competing techniques for different scenarios.

The paper is organized as follows: Section 2 provides a brief background on the different KF schemes considered. Next, the feeder used for testing and the simplified loss model adopted are presented in section 3. The implementation details of the KF based estimation techniques, as applied to the CPI problem, are given in section 4. In section 5, the results obtained in two different scenarios are presented and discussed, while the proposed KF technique is compared in section 6 with other published works dealing with CPI. The conclusions derived from the case studies considered are drawn in section 7.

2. Kalman filter background

In this section, three different Kalman filter formulations are succinctly reviewed, as applied to continuous-time, discrete-measurement nonlinear systems. In the discrete-time framework, the associated equations may be expressed as follows,

$$x_k = f(x_{k-1}, u_{k-1}) + w_k \tag{1}$$

$$z_k = g(x_k, u_k) + v_k \tag{2}$$

where x_k is the state vector at instant k, u_k the system input, and z_k the vector of available measurements. Gaussian processes are considered for the model and measurement noises, w_k and v_k, with covariance matrices Q and R, respectively.

The iterative processes of the Kalman filter schemes considered in this work, all of them involving prediction and correction stages, are described below.

2.1. Unscented Kalman Filter

At instant k, a cloud of $2L+1$ vectors, the so-called σ-points, is obtained from the previously estimated state vector, $\hat{x}_{k-1}$ (dimension L), and the covariance matrix of the state estimation error, P_{k-1}, as follows, [16]:

$$\begin{cases} x_{k-1}^0 = \hat{x}_{k-1} \\ x_{k-1}^i = \hat{x}_{k-1} + [\sqrt{(L+\lambda)P_{k-1}}]_i \\ x_{k-1}^{i+L} = \hat{x}_{k-1} - [\sqrt{(L+\lambda)P_{k-1}}]_{i+L} \\ i = 1, ..., L \end{cases} \tag{3}$$

$[\sqrt{(L+\lambda)P_{k-1}}]_i$ being the i^{th} column of the corresponding matrix, and λ a scaling factor calculated as follows:

$$\lambda = \alpha^2(L+\kappa) - L \tag{4}$$

where α and κ are two filter parameters to be tuned.

These σ-points are evaluated through equation (1), yielding $2L+1$ vectors, x_k^{i-}, from which the *a priori* estimations $\hat{x}_k^-$ and P_k^- are obtained:

$$\hat{x}_k^- = \sum_{i=0}^{2L} W_{mi} x_k^{i-} \tag{5}$$

$$P_k^- = \sum_{i=0}^{2L} W_{ci}(x_k^{i-} - \hat{x}_k^-)(x_k^{i-} - \hat{x}_k^-)^T + Q_k \tag{6}$$

where the weighting vectors W_m and W_c are calculated from:

$$\begin{cases} W_{m0} = \frac{\lambda}{L+\lambda} \\ W_{c0} = \frac{\lambda}{L+\lambda} + 1 - \alpha^2 + \beta \\ W_{mi} = W_{ci} = \frac{1}{2(L+\lambda)} \qquad i = 1, ..., 2L \end{cases} \tag{7}$$

β being another tunable parameter.

On the basis of the *a priori* information, the correction stage starts with the calculation of a new cloud of vectors, x_k^-, which are evaluated with the measurement function $g(\cdot)$ in equation (2), and weighted with the vectors W_m, yielding

$$\gamma_k^{i-} = g(x_k^{i-}, u_k) \qquad i = 0, ..., 2L \tag{8}$$

$$\hat{z}_k^- = \sum_{i=0}^{2L} W_{mi} \gamma_k^{i-} \tag{9}$$

Then, the covariance matrix of the measurement estimation error, P_{zk}^-, and the cross-covariance matrix of state and measurements, P_{xzk}^-, are obtained using the vector W_c as follows:

$$P_{zk}^- = \sum_{i=0}^{2L} W_{ci}(\gamma_k^{i-} - \hat{z}_k^-)(\gamma_k^{i-} - \hat{z}_k^-)^T + R_k \tag{10}$$

$$P_{xzk}^- = \sum_{i=0}^{2L} W_{ci}(x_k^{i-} - \hat{x}_k^-)(\gamma_k^{i-} - \hat{z}_k^-)^T \tag{11}$$

The correction stage concludes with the *a posteriori* predictions,

$$\hat{x_k} = \hat{x}_k^- + K_k(z_k - \hat{z}_k^-)^T \tag{12}$$

$$P_k = P_k^- - K_k P_{zk}^- K_k^T \tag{13}$$

which are based on the *a priori* predictions at instant k and the so-called Kalman gain, K_k, calculated from

$$K_k = P_{xzk}^-(P_{zk}^-)^{-1} \tag{14}$$

2.2. Cubature Kalman Filter

This KF formulation uses a set of $2L$ cubature points calculated from $\hat{x}_{k-1}$ and P_{k-1} through the following expressions, [17]:

$$S_{k-1}S_{k-1}^T = P_{k-1} \tag{15}$$

$$x_{k-1}^i = S_{k-1}\xi_i\sqrt{L} + \hat{x}_{k-1} \qquad i = 1, ..., 2L \tag{16}$$

where S is a positive-definite square root of matrix P (the Cholesky factorization of matrix P is customarily used), and ξ_i is the i^{th} cubature node, obtained as the intersection of the unit sphere and the $\mathbb{R}^L$ axis.

The state function $f(\cdot)$ in (1) is evaluated for the set of cubature points, yielding a set of $2L$ vectors x_k^{i-}, from which the *a priori* estimation is computed,

$$\hat{x}_k^- = \frac{1}{2L}\sum_{i=1}^{2L} x_k^{i-} \tag{17}$$

$$P_k^- = \frac{1}{2L}\sum_{i=1}^{2L} x_k^{i-}x_k^{i-T} - \hat{x}_k^-\hat{x}_k^{-T} + Q_k \tag{18}$$

For the correction stage, the covariance matrix P_k^- is factorized in order to calculate both the matrix S_k^-,

$$S_k^- S_k^{-T} = P_k^- \tag{19}$$

and a new set of $2L$ cubature points, x_k^-, at which function $g(\cdot)$ in (2) is evaluated to obtain γ_k^-.

Then the measurement estimation, $\hat{z}_k^-$, its covariance matrix, P_{zk}^-, and the cross-covariance matrix of state and measurements, P_{xzk}^-, are calculated as follows:

$$\hat{z}_k^- = \frac{1}{2L}\sum_{i=1}^{2L} \gamma_k^{i-} \tag{20}$$

$$P_{zk}^- = \frac{1}{2L}\sum_{i=1}^{2L} \gamma_k^{i-}\gamma_k^{i-T} - \hat{z}_k^-\hat{z}_k^{-T} + R_k \tag{21}$$

$$P_{xzk}^- = \frac{1}{2L}\sum_{i=1}^{2L} x_k^{i-}\gamma_k^{i-T} - \hat{x}_k^-\hat{z}_k^{-T} \tag{22}$$

The *a posteriori* predictions of the state vector, $\hat{x}_k$, and the covariance P_k are calculated with the Kalman gain using the same equations (14)-(13) as in the UKF algorithm.

2.3. Ensemble Kalman Filter

The EnKF, [18], is a Monte Carlo approximation of the original KF which has proven accurate enough in high-dimensional state-space problems. The ensemble is represented by an LxN matrix, N being the number of samples considered.

The ensemble is first propagated through the state and measurement functions,

$$\begin{cases} x_k^{i-} = f(x_{k-1}^i, u_{k-1}) \\ z_k^{i-} = g(x_k^i, u_k) \\ i = 1, ..., N \end{cases} \tag{23}$$

and then the mean values are calculated:

$$\overline{x_k^-} = \frac{1}{N}\sum_{i=1}^{N} x_k^{i-} \tag{24}$$

$$\overline{z_k} = \frac{1}{N}\sum_{i=1}^{N} z_k^{i-} \tag{25}$$

The EnKF correction stage is based on the calculation of the intermediate matrices

$$\overline{P_k}H_k^T = \frac{1}{N}\sum_{i=1}^{N}(x_k^{i-} - \overline{x_k^-})(z_k^{i-} - \overline{z_k})^T \tag{26}$$

$$H_k\overline{P_k}H_k^T = \frac{1}{N}\sum_{i=1}^{N}(z_k^{i-} - \overline{z_k})(z_k^{i-} - \overline{z_k})^T \tag{27}$$

which allow the Kalman gain and the updated values of each sample in the ensemble to be obtained:

$$K_k = \overline{P_k}H_k^T(H_k\overline{P_k}H_k^T + R)^{-1} \tag{28}$$

$$x_k^i = x_k^{i-} + K_k(z_k - z_k^{i-}) \qquad i = 1, ..., N \tag{29}$$

Finally, the corrected covariance matrix, P_k, is calculated as follows:

$$\overline{x_k} = \frac{1}{N}\sum_{i=1}^{N} x_k^i \tag{30}$$

$$P_k = \frac{1}{N}\sum_{i=1}^{N}(x_k^i - \overline{x_k})(x_k^i - \overline{x_k})^T \tag{31}$$

3. Problem statement and modeling

In order to assess the ability of the KF-based methodology to address the CPI problem, a typical distribution grid is considered comprising N_s single-phase and N_t three-phase customers, not necessarily balanced, resulting in $N_c = N_s + 3N_t$ total consumption curves in the network. In this work, it is assumed that the number of three-phase customers is 20% of the total. The proposed methodology, as presented in section 4, is easily extensible to other load connection arrangements, such as two-phase loads, still found in some areas. Figure 1 shows an example of a distribution network with 100 loads, used in the sequel to test the estimation techniques.

The energy consumption of each load i at a certain hour k, denoted by $E_{i,k}$, is obtained from [19], 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. In case of three-phase loads, a single-phase consumption is assigned to each phase. As customers with null consumption cannot be identified (they provide no information), the corresponding curves are removed from the raw data.

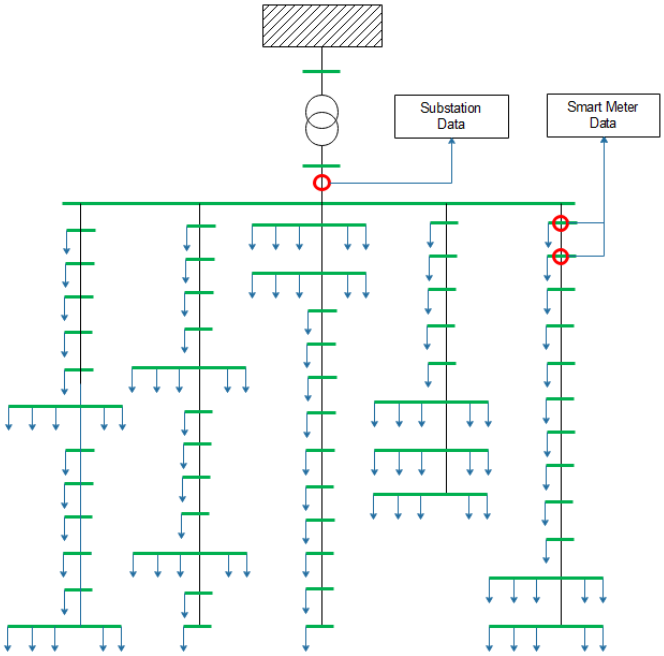

Figure 1: Single-line diagram of one of the test networks

To fully characterize the energy consumption of the customers, the reactive energy is also obtained from the raw data in [19]. Finally, the resulting hourly curves are randomly associated to a certain phase (a, b or c).

Along with the previous customer information, a typical distribution feeder topology is considered, each customer being associated to one of the grid nodes. Then, a load flow can be solved at each hour k in order to obtain the energy delivered by each phase of the MV/LV secondary substation, $E^a_{SS,k}$, $E^b_{SS,k}$ and $E^c_{SS,k}$, which is used by the proposed KF-based estimation technique as a measurement. This fully defines the distribution grid model involved in the estimation process.

4. Kalman Filter Implementation

The application of the different KF schemes described in Section 2 to the CPI problem is illustrated in the flowchart represented in Figure 2, as summarized in the sequel.

4.1. Parallel filtering

For every iteration of each tested technique, three independent KF-based estimators run in parallel, one for each phase p of the distribution grid, being the state vector composed of N_c variables, x^p_i, with $p = a, b, c$, so that $L = N_c$. Any three-phase customer is characterized by a set of three consecutive state variables, $\{x^p_i, x^p_{i+1}, x^p_{i+2}\}$,

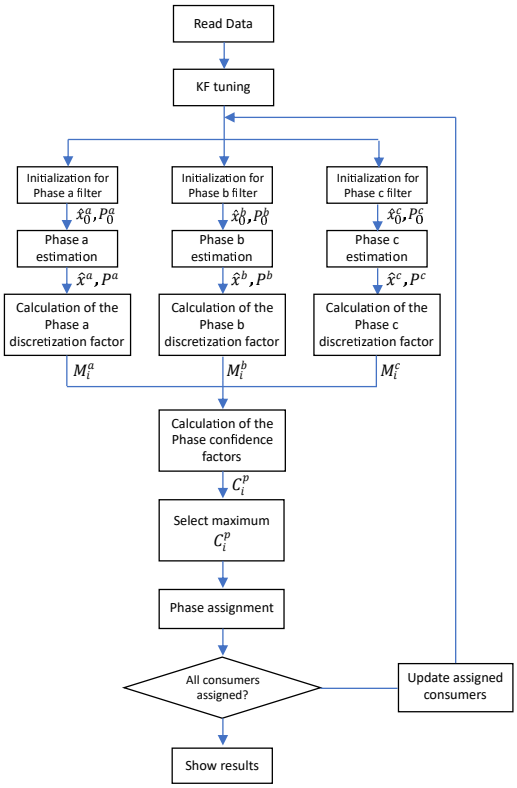

Figure 2: Flowchart of the CPI methodology

for each phase. Those variables indicate if the consumption i is associated to the corresponding phase p. For these static variables, the state function, $f(\cdot)$, in equation (1), is taken as a random walk, independent of the system input, u_k, yielding the following expression:

$$x^p_k = x^p_{k-1} + w^p_k \tag{32}$$

where the model noise, w^p_k, is considered to have the same covariance matrix, Q, for each phase.

The system input, u_k, used in the measurement function $g(\cdot)$ in equation (2), is determined by the hourly energy consumption of each customer, $u^T_k = E_1, E_2, ..., E_{N_c\,k}$. Regarding the network losses, as no measurements about the voltage and power factor of each load are available, a simplified loss model is adopted by the KF estimators, for which a given rated voltage and power factor $\cos\varphi_{i,k} = 1$ are assumed for all loads points. With these assumptions, an average hourly current, $I_{i,k}$ is calculated for each consumer i from the corresponding hourly energy consumption, $E_{i,k}$:

$$I_{i,k} = \frac{E_{i,k}}{T \cdot U} \tag{33}$$

where U is the voltage magnitude, and T is the energy integration period. In this work, as the energy measure-

ments are supposed to be obtained hourly, $T = 1h$.

Accordingly, the energy loss attributable to customer i is given by:

$$E_{i,k}^{loss} = I_{i,k}^2 \cdot T \cdot r \cdot l_i \tag{34}$$

where r is the conductor resistance per unit length and l_i is the electrical distance from each consumer to the secondary substation, both assumed to be known ($r = 0.223\Omega/km$, has been adopted in this paper).

The vector u_k also includes the variables x_i^p corresponding to customers which have been previously assigned to a phase, as described below.

The measurement of each filter, z_k^p, is taken as the energy delivered at each hour by the corresponding phase of the secondary substation, $z_k^p = E_{SS,k}^p$, so the measurement function $g(\cdot)$ in equation (2) reduces to:

$$z_k^p = \sum_{i=1}^{N_c} x^p{}_{i,k} \cdot (E_{i,k} + E_{i,k}^{loss}) + v_k^p \tag{35}$$

where the measurement noise, v_k^p, is considered to have the same covariance matrix, R, for each phase.

4.2. Initialization and tuning

As explained formerly, before the KFs are applied, a set of input data are gathered, including the hourly consumption of each customer and the energy delivered by the MV/LV transformer for the time window available. A separation is made between single-phase and three-phase loads. For the last ones, it is assumed that three energy measurements are available, one per phase, but the phase labels (a, b or c) are unknown. At the end of the estimation processes, the actual distribution of customers among the three phases is used to evaluate the performance of each estimation procedure.

The KF formulations require that an initial estimation is adopted, which is determined by the state vector $\hat{x}_0$ and the covariance matrix of the initial estimation error, P_0. For the UKF and CKF implementations, all state variables in the three phases are initialized as 0, assuming a complete lack of knowledge about their real values. In the case of EnKF, the samples are given random binary values as initial estimation, which enhances the convergence of the estimation algorithm. The covariance matrix, P_0 is considered as a diagonal matrix with $P_{ii} = 10$ in the three KF schemes.

The covariance matrices Q and R are defined as diagonal matrices, considering typical values of $Q_{ii} = 10^{-4}$ and $R_{ii} = 9 \cdot 10^{-4}$ respectively, the last one being equivalent to assuming a s.d. of 3% for the measurement errors.

For the estimation based on UKF, it is necessary to define the tunable parameters of the filters introduced in equations (4) and (7). A study is made in [20] over the influence of these parameters in the estimation process, concluding that $\alpha = 10^{-4}$, $\beta = 2$ and $\kappa = 3 - N_c$, are reasonable values for good estimation results.

Regarding the EnKF, the number of samples in the ensemble is taken as $N = 10 \cdot N_c$ so that it can suit the different sizes of the state vector.

4.3. Candidate selection and assignment

Once the three parallel estimators have provided a customer distribution for their corresponding phase, a single consumption is assigned and removed from the subsequent estimation processes. The candidate selection at each iteration of the CPI procedure is based on the estimated values of the state variables, $\hat{x}_i^p$, and the covariance of their estimation error, P_{ii}.

Forcing the state variables to be only 0 or 1, as required by the nature of the problem in hand, would involve equality constraints of the form $\hat{x} \cdot (\hat{x} - 1) = 0$. However, the application of the KF estimation algorithms to such non-convex model is prone to convergence problems. For this reason, the approach proposed in this work to enforce the binary character of the state variables relies on the confidence level of a given consumption being associated to a certain phase and not to the others. This is quantified through the so-called phase confidence factors, C_i^p, calculated with the probability density function (PDF) of the state variables, as follows.

First, the cumulative density function for $\hat{x}_i^p > 0.5$, denoted as M_i^p, is calculated as follows:

$$\begin{cases} M_i^p = \frac{1}{\sqrt{2 \cdot \pi \cdot P_{ii}}} \int_{0.5}^{\infty} e^{-\frac{(x - \hat{x}_i^p)^2}{2 \cdot P_{ii}}} dx \\ p = a, b, c \\ i = 1, ..., N_c \end{cases}$$

Since M_i^p provides information on the discrete value associated to the corresponding state variable, this coefficient is called phase discretization factor. Figure 3 includes a graphic representation of this factor, corresponding with the shaded area in the Gaussian density function, considering 0.5 as lower limit.

Then, for each energy consumption curve i, the phase confidence factor combine the information derived from the three phases of the corresponding load through the following expression:

$$\begin{cases} C_i^a = M_i^a \cdot (1 - M_i^b) \cdot (1 - M_i^c) \\ C_i^b = M_i^b \cdot (1 - M_i^c) \cdot (1 - M_i^a) \\ C_i^c = M_i^c \cdot (1 - M_i^a) \cdot (1 - M_i^b) \\ i = 1, ..., N_c \end{cases}$$

The maximum value of the phase confidence factors C_i^p is selected, which determines the consumption i with the highest probability of being associated to the phase p and not associated to the other phases.

Finally, once a consumption is selected using the previously defined phase confidence factors, the three variables associated with that customer are assigned integer values as follows:

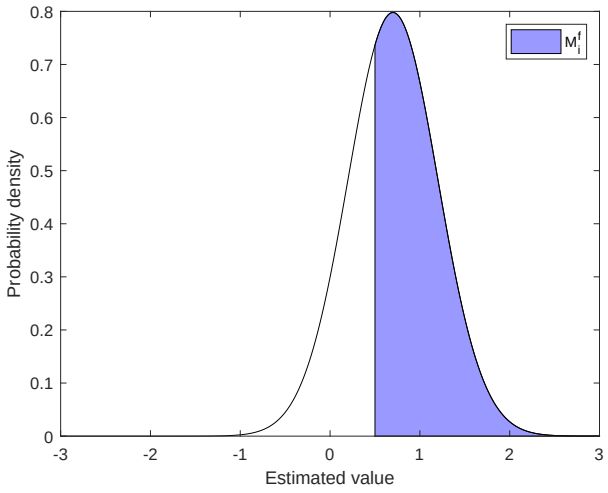

Figure 3: Representation of M_i^p

- If the consumption is related to a single-phase client, $\hat{x}_i^p = 1$ for the phase the client is connected to, and $\hat{x}_i^p = 0$ for the other two phases.
- For three-phase clients, assume without loss of generality that the consumption i is related to the phase a. Then, $\hat{x}_i^a = 1$, $\hat{x}_i^b = 0$ and $\hat{x}_i^c = 0$, as in the case of single-phase customers. Additionally, as two energy consumptions from a certain three-phase load cannot be associated to the same phase, the variables $\hat{x}_{i+1}^a$ and $\hat{x}_{i+2}^a$ are set to 0.

4.4. Update results

If all the loads have been assigned, then the estimation process ends. Otherwise, the state variables and the system input are updated.

For each filter, the state variables of the assigned consumptions are extracted from the state vector x and introduced as part of the system input, so that the number of unassigned consumers N_c is reduced accordingly. The estimation process is then repeated with the remaining consumers until $N_c = 0$, meaning that the final estimation result has been achieved.

5. Case studies

In this section, the proposed KF formulations are compared on several case studies, which can be grouped into four different scenarios.

By performing a series of preliminary tests in which the percentages of three-phase loads range from 10 to 30%, no significant differences have been observed in the performance of the proposed methodology. For this reason, as stated above, 20% of three-phase customers has been assumed in all scenarios.

5.1. Scenario I: Original measurements

In this scenario, the actual consumption is used for the KF-based estimation, assuming that all the measurements are correctly obtained.

The performance of UKF and CKF with increasing number of loads is shown in Tables 1 and 2, respectively. Note that, while both estimators correctly assign 100% of customers for $N_c = 50$, the behavior of the UKF deteriorates faster as the number of loads increases.

Consumption curves	Correct assignments	Wrong assignments	Percentage
50	50	0	100%
100	95	5	95%
200	181	19	90.5%
300	216	84	72%
400	233	167	58.25%

Table 1: Estimation results for the UKF. Scenario I

Consumption curves	Correct assignments	Wrong assignments	Percentage
50	50	0	100%
100	100	0	100%
200	195	5	97.5%
300	262	38	87.33%
400	313	87	78.25%

Table 2: Estimation results for the CKF. Scenario I

Consumption curves	Correct assignments	Wrong assignments	Percentage
50	50	0	100%
100	100	0	100%
200	200	0	100%
300	283	17	94.33%
400	346	54	86.5%

Table 3: Estimation results for the EnKF. Scenario I

As shown in Table 3, the EnKF response to increasing system sizes is better than that of the other formulations, which confirms the expected behavior of this KF scheme for high-dimensional problems.

The convergence of the estimation processes is not only determined by the value of $\hat{x}_i^p$, but also by the covariance of the estimation error for each state variable, P_{ii}. An illustrative example is shown in Figure 4, representing the PDFs obtained for the three state variables of a certain single-phase consumer, given the values of $\hat{x}_i^p$ and P_{ii} from the KF-based estimation process. The phase discretization factors, M_i^p, corresponding to the shaded areas in

the graphics, and their numeric values are included in the respective legends. In this particular case the computed values of the phase confidence factors are $C_i^a = 0.014$, $C_i^b = 0.711$ and $C_i^c = 0.001$, meaning a confidence of 0.711 that the consumer is associated to phase b. This value would be compared with those of the rest of the clients in order to obtain the selected candidate in the corresponding iteration of the proposed methodology.

Finally, in order to illustrate the evolution of the phase confidence factor, C_i^p, Figure 5 represents this coefficient in descending order at different stages of the estimation process for 100 loads (i.e., 300 C_i^p factors), using the EnKF formulation. It can be noticed that the maximum of the confidence factors, used to select the next candidate at the corresponding iteration, is close to 1 in all cases.

In the top graph of Figure 5, when the whole set of 100 loads is still unassigned, a large number of phase confidence factors remain with a small value ($C_i^p = 0.125$), in accordance to the common initial value $M_i^p = 0.5$ adopted for the three-phase discretization factors, which means a complete lack of information, at this early stage of the iterative process, on the phase to which those consumption curves should be associated. Then, as more loads are assigned, the coefficients of the remaining loads more clearly show a trend towards 1 or 0 (about one third tends to 1 whereas the remaining two thirds tend to 0).

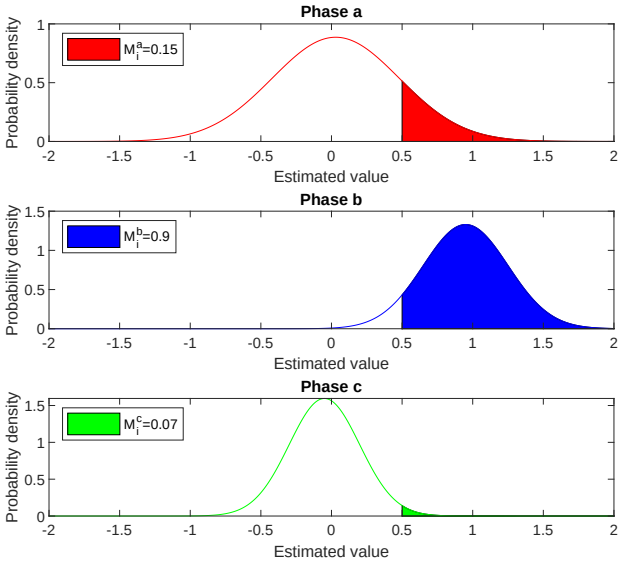

Figure 4: Probability density functions for a sample consumer

5.2. Scenario II: Analysis of the required amount of data

For the results shown above, hourly smart meter readings for 20 days are used, leading to 480 total energy measurements for each consumer. This section analyzes the quality of the estimation as the number of available measurements decreases. Table 4 summarizes the performance

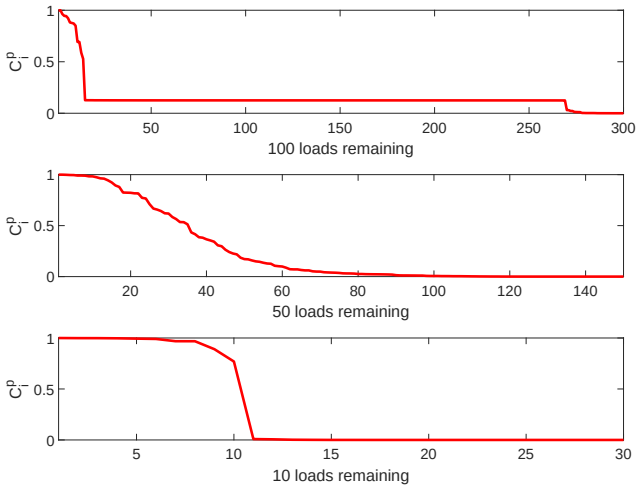

Figure 5: Evolution of the phase confidence factor at different stages

of the three estimators for the case in which $N_c = 200$ consumption curves ($N_s = 116$ single-phase and $N_t = 28$ three-phase customers).

Available data	UKF Hit Rate (%)	CKF Hit Rate (%)	EnKF Hit Rate (%)
480	90.5	97.5	100
400	90.5	97.5	100
300	90	96	100
200	85	90	95
100	67	70	72

Table 4: Estimation results for different KF formulations. Scenario II

Those results suggest a deterioration of the performance of the KF-based estimation techniques when the number of available measurements is lower than 200 (around 8 days) in this particular scenario. Further tests with feeders comprising different numbers of customers show that the required number of measurement snapshots increases with the number of loads, as intuitively expected. For instance, for $N_c = 100$ (one half of that shown in Table 4), the success rate of the three estimators does not deteriorate substantially, even when only 100 energy measurements (around 4 days) are considered for each load.

5.3. Scenario III: Noisy measurements

The performance of the different KF formulations is evaluated in a realistic scenario where errors in the measurements are considered. As the objective of these case studies is to determine the robustness of the KF schemes against measurement errors, a relatively low number of loads is considered, namely $N_s = 58$ single-phase and $N_t = 14$ three-phase clients, which leads to a total of $N_c = 100$ consumption curves.

Gaussian noise is artificially added to each measurement after the load flow is computed. The rates of correct phase-to-customer assignments for increasing measurement errors are summarized in Table 5 for the three estimators.

Noise level	UKF Hit Rate (%)	CKF Hit Rate (%)	EnKF Hit Rate (%)
1%	95	100	100
2%	92	96	100
3%	89	93	98
5%	83	89	92

Table 5: Estimation results for different KF formulations. Scenario III

In light of those results, it can be concluded, also as expected, that the number of correct assignments decreases with increasing measurement noise, for every KF formulation, being the robustness of the CKF and EnKF similar, superior in any case to that of the UKF formulation. Nevertheless, all formulations show acceptable results when typical noise levels are considered in the measurements.

5.4. Scenario IV: Model errors

In the proposed implementation of the KF for the CPI problem, a simplified loss model is considered for which the value of the conductor resistance per unit length, r, is required as per equation (34). In this scenario, the performance of the different KF formulations is evaluated when errors in r are considered.

For the same number of customers as in Scenario III, Table 6 summarizes the rates of correct phase-to-customer assignments for errors in r ranging from 5 to 20%.

Error in r	UKF Hit Rate (%)	CKF Hit Rate (%)	EnKF Hit Rate (%)
5%	95	100	100
10%	91	95	98
15%	86	91	95
20%	80	84	89

Table 6: Estimation results for different KF formulations. Scenario IV

It can be concluded that the results remain acceptable, at least for the EnKF, when the assumed resistance error does not exceed 10%.

6. Comparison with existing methods

Considering the nature of the information on which the proposed KF methods are based, a comparison can be easily made with the performance of the methods proposed in [7], where a LASSO-based technique is applied to the CPI problem using exclusively energy consumption curves obtained from smart meters, and in [8], where PCA is considered for the same purpose. Different scenarios are considered for the comparison, with increasing number of consumers and 1% s.d. in the measurement errors for all cases.

The rate of correct assignments for each method in the different cases studied are summarized in Table 7, where it is observed that:

- The performance of the PCA method is similar to that of the UKF in all the studied scenarios.
- The LASSO-based method has proven to be more sensitive to the number of loads than the CKF and the EnKF formulations, being the success rates of the three techniques very close for $N_c \leq 200$.
- The results obtained by the CKF and the EnKF schemes are substantially better when the number of loads increases, as can be noticed particularly for $N_c = 400$.

Number of clients	PCA	LASSO	UKF	CKF	EnKF
50	100	100	100	100	100
100	100	100	95	100	100
200	92	100	90.5	97	100
300	72.67	87.67	71	86.33	94
400	60.25	71.75	57.75	78.25	86.5

Table 7: Rate of correct assignments for the PCA, the LASSO-based and the proposed KF-based methods

7. Conclusions

In this paper, a technique based on Kalman filtering and smart meters information is presented to identify the electrical phase which individual loads are connected to in distribution grids. For this purpose, the performance of three KF schemes, UKF, CKF and EnKF, is tested and compared.

The proposed estimation algorithm iteratively selects the customer with the highest probability to be connected to a certain phase, based on the estimated value of the corresponding state variable and the covariance of the estimation error. This way of handling binary variables prevents the computational problems potentially arising by the enforcement of the customary equality constraints, and might find application in other binary-constrained problems.

Four separate scenarios have been considered to test the accuracy and robustness of the KF formulation for feeders with increasing number of loads and increasing measurement and model errors. The amount of measurements required for a good performance of the algorithm is also analyzed. From the results obtained, the following conclusions are drawn:

- The UKF shows a poor behavior as the number of loads increases, achieving unacceptable success rates ($< 60\%$) when 400 customers are considered. Additionally, the robustness of this formulation in the presence of wrong measurements is weak.
- The performance of the CKF is better than that of the UKF, showing a lower sensitivity to the state vector size and wrong data.
- The EnKF has proven to be the best of the three KF schemes considered for the CPI problem, both in terms of success rates and sensitivity to noisy measurements.
- As expected, the performance of the KF-based schemes is affected by the number of energy measurements available, relative to the number of loads. The more loads in the same feeder, the more measurements are needed to correctly ascertain the phase connection.

A comparison of the CKF and EnKF with published methods, based on PCA and LASSO, has shown that the latter provide similar results to those of the KF only when a reduced number of loads is considered.

Further research efforts will be devoted to redesign the proposed KF-based procedure, so that it can take advantage of additional electrical quantities, such as P, Q and V, potentially available in the context of advanced DMS.

Acknowledgement

The authors thank the Ministry of Education and Professional Training of Spain (grant FPU17/06380) and the project Pastora (PI-1897/12/2019) for the financial support.

References

[1] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in: 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, 2012, pp. 1-7, Oct. 2012.

[2] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase LV distribution network," in: 2012 22nd Australasian Universities Power Engineering Conference (AUPEC), Bali, 2012, pp. 1-7, Sept. 2012.

[3] L. Blakely, M. J. Reno and W. Feng, "Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries," in: 2019 IEEE Power and Energy Conference at Illinois (PECI), Champaign, IL, USA, 2019, pp. 1-7, March 2019.

[4] C.S. Chen, T. T. Ku and C. H. Lin, "Design of phase identification system to support three-phase loading balance of distribution feeders," in: IEEE Trans. Industry Applications, vol.48, no.1, pp.191-198, Jan./Feb.2012.

[5] K. Caird, "Meter phase identification," U.S. Patent Application 20100164473, Patent No. 12/345702, Jan. 2010.

[6] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavloski, S.Hoy and J.R. Kalagnam, "Phase identification in smart grids," in: Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on, pp.25-30, Oct. 2011.

[7] X. Tang and J. V. Milanovic, "Phase Identification of LV Distribution Network with Smart Meter Data," in: 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, 2018, pp. 1-5, Dec. 2018.

[8] Satya Jayadev P, A. Rajeswaran, N. P. Bhatt and R. Pasumarthy, "A novel approach for phase identification in smart grids using Graph Theory and Principal Component Analysis," in: 2016 American Control Conference (ACC), Boston, MA, 2016, pp. 5026-5031.

[9] M. Xu, R. Li and F. Li, "Phase Identification With Incomplete Data," in: IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 2777-2785, July 2018.

[10] Pengxiang Ren, Hanoch Lev-Ari, and Ali Abur, "Robust Continuous-Discrete Extended Kalman Filter for Estimating Machine States with Model Uncertainties," in: Power Systems Computation Conference (PSCC), June 2016.

[11] Zhou Ning, Meng Da, Huang Zhenyu, G. Welch, "Dynamic state estimation of a synchronous machine using PMU data: A comparative study," in: IEEE Power & Energy Society General Meeting, Jan. 2015.

[12] J. Qi, K. Sun, J. Wang, H. Liu, "Dynamic state estimation for multi-machine power system by unscented Kalman filter with enhanced numerical stability," in: IEEE Transactions Smart Grid, March 2016.

[13] A. Rouhani and A. Abur, "Constrained Iterated Unscented Kalman Filter for Dynamic State and Parameter Estimation," in: IEEE Transactions on Power Systems, vol. 33, no. 99, pp. 1-1. doi: 10.1109/TPWRS.2017.2764005, May 2018.

[14] M.A. González-Cagigal, J.A. Rosendo-Macías, A. Gómez-Expósito, "Parameter estimation of fully regulated synchronous generators using Unscented Kalman Filters", in: Electric Power Systems Research, Volume 168, 2019, Pages 210-217, November, 2018.

[15] Y. Wehbe, L. Fan, "UKF based estimation of synchronous generator electromechanical parameters from phasor measurements," in: North American Power Symposium (NAPS), Sept. 2012.

[16] D. Simon, "Optimal State Estimation: Kalman, H Infinity and Nonlinear Approaches". ISBN: 13 978-0-471-70858-2 .

[17] H. Pesonen and R. Piché, "Cubature-based Kalman filters for positioning", in: Positioning Navigation and Communication (WPNC), 2010 7th Workshop on , vol., no., pp.45,49, 11-12, March 2010.

[18] S. Afrasiabi, M. Afrasiabi, M. Rastegar, M. Mohammadi, B. Parang and F. Ferdowsi, "Ensemble Kalman Filter based Dynamic State Estimation of PMSG-based Wind Turbine," 2019 IEEE Texas Power and Energy Conference (TPEC), College Station, TX, USA, 2019, pp. 1-4.

[19] Koirala Arpan, Arboleya Pablo, Mohamed Bassam and Suárez-Ramón, Lucía. "Non-Synthetic European Low Voltage Test System," in: International Journal of Electrical Power & Energy Systems. 118. 10.1016, May 2019.

[20] J.J. Hyun, L. Hyung-Chul, "Analysis of scaling parameters of the batch unscented transformation for precision orbit determination using satellite laser rangingdata", in: J. Astron. Space Sci. 28 (2011) 183–192.