Phase Identification in Low Voltage Smart Grids Using Smart Meter Data

João L. T. Santos

***Abstract* — Low voltage distribution grid characterization often lacks information on customer's phase connectivity. This leads to obvious ineffectiveness in maintaining phase-load balance, which, in turn, may cause several operation inefficiencies such as increased energy losses and unnecessary voltage imbalances. Yet, with the deployment of smart metering and the consequent availability of energy consumption data of pre-defined time-resolution, phase connectivity information might be possible to estimate, if data on per-phase aggregate energy measurements are available at substation sites with the same time-resolution. In this thesis, a set of data analytics tutorial approaches to identify the underlying customer phase-connectivity from time series of energy consumption and their aggregated per-phase energy measurements were studied. Based on the study, a new method, which applies Multivariate Linear Regression, is then implemented and compared with state-of-the-art methods based on Principal Component Analysis. Comparisons were carried out with experimentation (*i*) in laboratorial conditions where aggregated per-phase energy measurements data is built to replicate typical grid losses, random noise, energy theft, and clock skew and also synchronization errors, but also (*ii*) with real-world data provided for a specific location by Portugal's incumbent EDP Distribuição. Results have shown that the new Multivariate Linear Regression method consistently presented better performance than the state-of-the-art methods, both in extreme laboratorial and near-real world conditions.**

***Index Terms* — Phase Identification, Low Voltage Smart Grids, Smart Meters, Multivariate Linear Regression, Principal Component Analysist.**

I. Introduction

PHASE identification is a critical input to the grander problem of phase load balancing [1]. As electricity is usually generated and distributed as three-phases separated by 120º AC voltage, households mostly draw from a single phase, and maintaining phase load balance in substation transformers is paramount to achieve network efficiency and prolonging the life time of assets [2], [3].

As consumers become more technology and environmentally conscious, power utility companies face the challenge of managing revenue recession while meeting the demands of their customers in a progressively more complex and dynamic distribution network [4].

In fact, rapid growth in Distributed Energy Resources (DERs), primarily solar, and plug in devices, such as electrical vehicles, due to indorsement by governments through lighter taxation, is requiring a more active management of the distribution network as an answer to more frequent network configuration changes [5]–[7].

Utilities are responding to these challenges by seeking increased efficiency while innovating, namely by investing heavily into smart grids which allow the implementation of analytics solutions to augment Automated Metering Infrastructure (AMI) productivity. Actually, it is forecasted that global investment in analytics solutions and integration services with this goal will amount to $10.1 billion through 2021 [8].

However, despite these investments, many important applications for network control and optimization such as 3-phase power flow optimization, volt-VAR control, distribution network state estimation, reconfiguration and restoration and load balancing, still rely on the network connectivity model and phase connectivity being known [9]. While the connectivity model is mostly reliable, phase connectivity information is often erroneous or missing. This is due to repairs, maintenance and common phase balancing projects that do not update phase connectivity information [2], [10].

Whereas distribution grid configuration and phase load balancing are key to reduce power loss and integrating DERs, incorrectly classifying the phase of a household or cable may lead to further unbalancing and possible overloads, which may lead to higher copper losses, voltage drops or equipment damage and consequent service interruption [2], [11], [12].

Historically, solving the phase identification problem relied on hardware-based methods. These however, require additional equipment or workforce to operate it, which can became a costly solution [13]. On the other hand, recent studies have taken a data analytical approach to solve the phase identification problem. Several machine-learning algorithms have been proposed, nevertheless the proposed methods tend to be computationally intensive and complex to implement. Thus, this paper seeks to present a novel and simpler method for phase identification, utilizing Multivariate Linear Regression (MLR) while comparing its performance to the state-of-the-art method proposed in [1], which utilizes Principal Component Analysis (PCA).

The paper is organized as follows. In Section II, the predictive models implemented in this paper are presented. In Section III the methodology to generate simulation data for the phase identification problem is explained. In Section IV, the application of the methodology is illustrated with simulated data and the quality of the results obtained is discussed. In Section V, the paper is concluded.

João L. T. Santos is with Deloitte Consultores SA., 1070-100 Lisbon Portugal (e-mail: joaotsantos@deloitte.pt).

II. Predictive Models for Phase Identification

In this section, the proposed method to infer customer phase connectivity is described and detail on the benchmark method for comparison (PCA) is given.

A. Multivariate Linear Regression (MLR)

In statistics, linear regression is used when considering the linear relationship between one or more scalar dependent (or response) variables y and one or more independent (or explanatory) variables x [14].

Its application is often categorized in two comprehensive groups:

1. Prediction or forecasting: utilizing the linear regression to fit a model through a dataset and then predict the dependent variable for a new input set of x's;
2. Quantifying relationship between variables: identify which subsets of x's contribute to explaining y, and how strongly.

Different linear regression applications are distinguished based on the number of dependent and independent variables, which determines the model name:

1. Simple Linear Regression: One y and one x, a single independent variable is used to predict the behavior of the dependent variable;
2. Multiple Linear Regression: One y and multiple x's, using more than one explanatory variable to explain the response variable;
3. Multivariate Linear Regression (also referred to as Multivariate Multiple Linear Regression): Multiple y's and multiple x's, relationship between different explanatory variables and possibly correlated independent variables to measure the influence of each of the dependent variables on each response variable.

The basic model for a Linear Regression is given by:

$$y_i = \beta_0 1 + \beta_1 x_1 + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \qquad (1)$$

Where β_i represents the parameter vector and β_0 is the constant offset term, ε_i corresponds to the error or noise and $x_i^T\beta$ is the inner product of vectors x_i and β.

Specifically, MLR is the implementation that best fits the problem discussed in this paper. For every set of x's there is a corresponding set of y's measured, related by different parameters, which can be expressed in matrix form by:

$$Y = XB + E \qquad (2)$$

Where the n dependent values measured for the p independent variables are given by:

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} = \begin{matrix} y_1' \\ \vdots \\ y_n' \end{matrix} \qquad (3)$$

And the dependent variables are stacks in the X matrix as follows:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1q} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nq} \end{pmatrix} \qquad (4)$$

Summarizing the model dimensions, Y is ($n \times p$), X is ($n \times (q+1)$) and B is ($q+1$).

The employment of MLR is based on some assumptions that lead to good estimates:

1. $E(\epsilon_i) = 0$, the expected value for the error is zero;
2. $cov(y_i) = \Sigma$, each row of Y has the same covariance matrix;
3. $cov(y_i, y_j) = 0$, rows of Y are uncorrelated with each other

However, these assumptions will be challenged in the implementation of the model to solve the phase connectivity problem when noise is added.

In order to find B, Ordinary Least Squares (OLS) approach is one of the more common approaches for fitting the linear regression model. Considered one of the simplest methods and computationally straightforward, OLS minimizes the sum of the squared residuals, and the formula is given by:

$$B = (X^T X)^{-1} X^T Y \qquad (5)$$

B. Principal Component Analysis (PCA)

In order to establish a basis for performance comparison, a basic implementation of PCA was also developed, following the work of Satya *et al*. [1].

PCA is widely spread as a tool for multivariate analysis. It is a statistical procedure that aims to obtain linearly uncorrelated variables, nominated principal components, from a dataset of observations of possibly correlated data by means of an orthogonal transformation. PCA is applied by eigenvalue decomposition of a covariance matrix or Singular Value Decomposition (SVD) of a data matrix. It is considered to be the simplest of multivariate analysis based on eigenvectors.

The objective of network model identification with PCA is to obtain the true data subspace and constrained subspaces from a data matrix Z, where Z is a ($n \times m$) matrix with n number of nodes or meters, including aggregated measures, and m number of measurements or samples per node.

The n variables are linearly related, with p linear relationships, given by:

$$CZ = 0 \qquad (6)$$

Where C is the ($p \times n$) constraint matrix.

These subspaces are obtained from the eigenvectors of the covariance matrix $S_Z = ZZ^T$, which can be attained by using the SVD of Z, such that:

$$SVD(Z) = U_1 S_1 V_1^T + U_2 S_2 V_2^T \qquad (7)$$

Where U_1 and U_2 are the set of orthogonal eigenvectors corresponding to the $(n-p)$ largest and p smallest eigenvectors of S_z respectively, with p dependent variables and $(n-p)$ independent variables, and S_1 and S_2 are diagonal matrixes with the singular values of Z.

In [15], it has been shown that the subspace S_R covered by the rows of U_2^T and C are equivalent:

$$S_R(U_2^T) \sim S_R(C) \qquad (8)$$

Therefore, by replacing C in (7) the following relationship is obtained:

$$U_2^T Z = 0 \quad (9)$$

However, given that the constraint matrix suffers from rotational ambiguity, the estimated constrained matrix $\hat{C}$ is not unique and may not be the correct solution that represents the physical interpretation of the problem:

$$U_2^T Z = \hat{C} Z = Q \hat{C} Z = 0 \quad (10)$$

Where Q is a non-singular matrix.

To achieve a unique solution, a regression model can be obtained by subdividing variables into dependent and independent variables:

$$Z = \begin{bmatrix} Z_d \\ Z_i \end{bmatrix} \quad (11)$$

Where Z_d represents the first rows of the Z matrix with the p dependent variables and Z_i the $(n-p)$ last rows with the independent variables.

Also, the constraint matrix $\hat{C}$ can be partitioned as well into a $(n_d \times n_d)$–dimension $\hat{C}_d$ matrix and a $(n_d \times n_i)$–dimension $\hat{C}_i$ matrix:

$$\hat{C} = \begin{bmatrix} \hat{C}_d \\ \hat{C}_i \end{bmatrix} \quad (12)$$

Consequently, from (10) it is possible to obtain:

$$\hat{C}_d Z_d + \hat{C}_i Z_i = 0 \quad (13)$$

Finally, since U_{2d} is of full rank, (13) can be expressed in terms of the regression matrix relating the dependent and independent variables so that:

$$Z_d = -(\hat{C}_d)^{-1} \hat{C}_i Z_i = \hat{R} Z_i \quad (14)$$

Where $\hat{R}$ is the $(n_d \times n_i)$–dimensional regression matrix, proven to be unique in [15].

In conclusion, the regression matrix using PCA is given by:

$$\hat{R} = -(\hat{C}_d)^{-1} \hat{C}_i \quad (15)$$

C. Time complexity of the algorithms

Although accuracy of the algorithms to correctly identify customer-to-phase connectivity is the principal performance measure employed in this work, it is relevant to refer to the time complexity of the algorithms.

In computer science, time complexity, usually presented with the O-notation, is a formal measure to estimate the time it takes for the algorithm to run.

Considering n as the number of nodes and m as the number of measurements per node, when applying the MLR algorithm it takes:

- $O(n^2 m)$ to multiply $X^T X$
- $O(nm)$ to multiply $X^T Y$
- $O(n^3)$ to compute the Cholesky factorization of $X^T X$ and use that to compute $(X^T X)^{-1} X^T Y$

Since in most of the simulations $m > n$, $O(n^2 m)$ asymptotically dominates over other computations and therefore it is considered the time complexity for applying OLS with MLR.

Complementary, in [1], the time complexity of the PCA algorithm is demonstrated to be $O(nm^2)$, due to the Singular Value Decomposition (SVD) of Z which is the most expensive step.

Thus, taking into consideration that usually the number of measurements m is greater than the number of customers n, although very similar in complexity, the MLR algorithm is proven to be better performant in an ideal implementation.

III. Methodology

In order to accomplish the scope of this work, the algorithms in analysis were implemented in R Studio [16] programming language, in a computer with Windows 10 – 64bit , CPU @ 2.30GHz and 12,0GB RAM.

Firstly, the program starts by importing consumer daily profiles' input data from a text file into the application environment. Necessary data cleansing is performed and a data table with daily consumer profiles is built.

Secondly, a phase is randomly attributed to each client, following a uniform distribution, and aggregated phase totals are calculated, simulating secondary substation readings.

Afterwards, different types of errors or noise are introduced to the aggregated phase totals, and true customer phase is hidden.

Subsequently, true customer smart meter readings and erroneous data simulating secondary substation phase totals are then fed to both MLR and PCA algorithms which compute the customers' attributed phase.

Finally, algorithm accuracy is then calculated based on whether the algorithm correctly predicts customer-to-phase allocation.

A. Smart meter and secondary substation load measurements

For the development of this paper, centered on the time-series of energy measurements from both consumer smart meters and secondary substation readings, a sample of daily consumer load profiles has been provided by an energy company for an undisclosed location.

Ideally, in real world situations, input data will be supplied including secondary substation readings with phase totals aggregated per phase. However, since this work is developed under laboratorial conditions and because known information does not include secondary substation readings, these need to be simulated following the methodology introduced in the previous section and detailed subsequently.

Input data consisted of 1623 daily consumer load profiles, each with a total of 96 readings, measured every 15 minutes. The time series of load measurements is in kW. Where information was unavailable, it was considered null.

In order to create consumer profiles which spawn more than one day, daily load profiles were grouped together, depending on the number of customers and necessary number of days to achieve the target number of measurements per number of clients' ratio. Fig. 1 represents the load diagrams for a sample of two customers, spawning over three consecutive days.

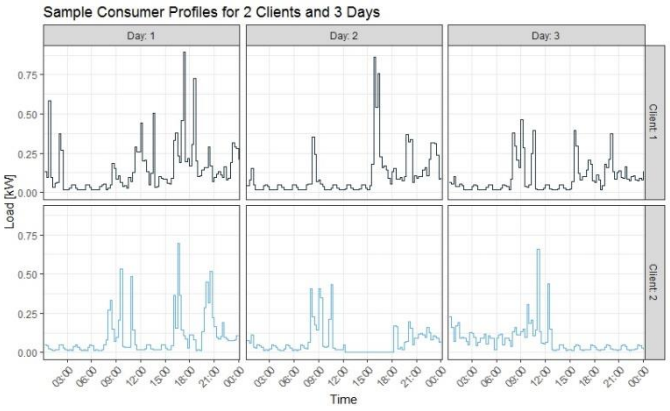

Fig. 1. Sample daily load profiles for 2 clients and 3 days. It is possible to observe in Client 2's second day of readings, between 12:00 and 18:00, an example of the missing data which may arise due to mechanical faults, human error, fraudulent behavior, instrument error or changes in system behavior [17].

Subsequently, customer phase was randomly allocated following a uniform distribution and load profiles were aggregated according to their new allocated phase. Fig. 2 simulates the readings from a secondary substation.

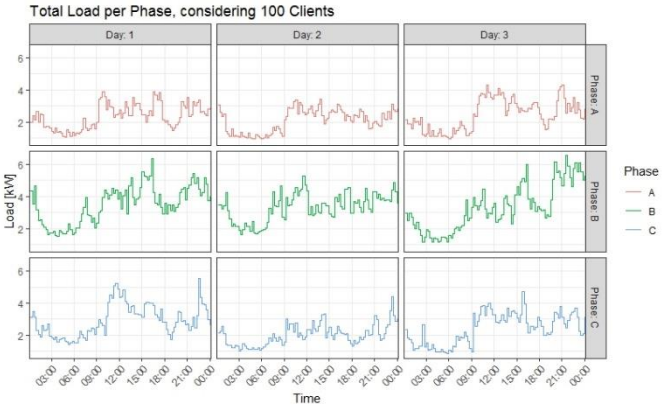

Fig. 2. Simulated secondary substations load measurements for each phase.

This representation, where the readings on the secondary substation are exactly equal to the sum of the readings on the smart meters, would only be accurate if there were no errors and no noise. However, in practice, such errors are unavoidable and thus the following section explores different types of errors considered in the simulations.

B. Noise modeling

In order to test the robustness of both algorithms under non-ideal situations, we introduced several types of errors:

1) *Meter accuracy class*: Electricity smart meters inherently have an accuracy class, result of its design, build quality and other factors. Understandably, a higher quality measuring meter will provide better accuracy but have significantly increasing costs for the utilities company. Thus, standards are defined to stipulate the minimum accuracy ratings required for smart meters [18].

ANSI C12.20 states that for smart/electronic meters must have at the very least 0.5 accuracy class, while IEC/AS Standard 62053 describes the requirements for 0.5, 1 and 2 accuracy classes. In this work, 0.5 accuracy class meters were considered as a reference for the typical error which means readings must be in the range of ± 0.5% of the true value.

This error may be approximately modeled by multiplying every reading with a random value following a Gaussian distribution with mean 1 and standard deviation 1/3 of meter accuracy, such that 99.7% of simulated errors fall within the defined 0.5 accuracy class.

2) *Clock asynchronism:* Next, two types of clock errors were introduced, commonly modelled together but, in this exercise, simulated independently.

Firstly, clock asynchronism is a result of clocking the load at different points in time and thus the measurement of total load for a given time is not exactly the sum of smart meter readings for that time interval. Unlike the meter accuracy error, clock asynchronism does not change with time.

In an effort to increase efficiency in existing smart grid infrastructure, utilities are progressively more dependent on high quality data that must be synchronized with very high accuracy for control and protection as well as data analytics solutions. Multiple applications such as measurement systems, fault locators or protection relays require microsecond precision from substation readings. Synchronous sampling is critical as it can introduce errors in solutions but for customer end-points requirements are not so strict and thus small synchronization errors can influence phase identification models [19].

Following V. Arya *et al*. [2] implementation, to simulate clock asynchronism, each meter is made erroneous by adding a random Gaussian walk. Instead of clocking the load after every Δt units, the k^{th} measurement clocks the load for the time interval $[T_{K-1}, T_K]$ where $T_K = T_{K-1} + N(\mu = \Delta t, \sigma = f\Delta t), f \in [0, 2.23]\%$. In summary, in this simulation, all clocks considered must have a maximum (3σ) of ±1 min asynchronism which, taking into account readings are measured every 15 minutes, corresponds to 6.67%.

3) *Clock skew:* Introducing the second type of clock error, clock skew occurs when each smart meter's internal clock runs at a different frequency from that of the true clock which, in this smart grid application may be considered as the substation clock.

Usually, a single clock signal is used to synchronize all clock frequencies. However, one disadvantage associated with this technique is that each microprocessor in smart meters may receive the signal at different points in the chip. Moreover, several factors may contribute for causing clock skew such as electromagnetic propagation delays, buffer delays in the distribution network, differences in temperature, variations in the manufacturing process, power supply variations and different load capacitance [20].

In this simulation, in order to compute the frequency of each meter's clock in comparison to the substation clock, a random shift in frequency is introduced following a Gaussian distribution so that it lies in the interval $[-f\Delta t, f\Delta t], f \in [0, 30]\%$.

Although a maximum shift in frequency of 30% is considered as the base case, this is a very high skew error since the skew error for a real clock usually lies in the order of milliseconds [2].

4) *Copper losses:* Low voltage distribution networks enable the transmission of electric energy from secondary substations to customers in independent households through large and complex networks. These networks consist of not only overhead lines or buried cables but also other equipment such

as transformers. As previously stated, the hard fact is that there are always losses in the network and thus the generated electric energy does not match with the total energy supplied to consumers. Losses may be classified as technical or commercial losses [21].

In this segment, technical copper losses were introduced which can be due to energy dissipated in the conductors and equipment used for transmission, transformation or distribution. In the European Union, is it estimated that around 4% of total generated energy is wasted due to distribution losses [22].

Copper losses, due to resistance along the wirelines or internal wiring within the transformers, scale with current squared time resistance (I2R) and the majority of distribution line losses occur within the primary and secondary distribution lines.

In this simulation, the base case is considered to have copper losses in the $[2\%, 10\%]$ interval, varying quadratically with load.

5) *Missing Clients:* Another type of network losses may be due to commercial losses. In low voltage distribution networks, customers have to pay their electricity bills according to their unit consumption and their particular needs, depending upon the contracted tariff. Specifically in smart grids, the devices used to measure power consumption for billing purposes and network control are smart meters.

Although smart meters are harder to tamper with than electromechanical KWh meters, billions of dollars are lost every year to electricity theft. There are multiple ways of sabotaging energy measurement such as unauthorized extensions of loads, tampering the meter readings by mechanical jerks, placement of powerful magnets or disturbing the disc rotation with foreign matters, stopping the meters by remote control, changing of terminal wiring, changing current transformer ratio or even some involuntary actions such as improper testing and calibration of meters [23].

While in developed countries secure networks experience only around 1-3% electricity theft, developing countries have been shown to have much higher theft percentages [24].

In our simulation, a sample of 5 random customer load profiles were added to the substation totals in order to simulate energy theft. Considering an example of 100 clients, this corresponds to 5%.

Predictively, introducing missing clients' error, will have a great impact on phase identification algorithms because it introduces a variation in substation totals that is in no way dependent on given customer readings.

C. Model implementation and performance measures

This section explains how the MLR and PCA algorithms were applied to the problem of phase identification and implemented in RStudio.

Firstly, the model for MLR is presented. The output is matrix X, with dimension 3 by n customers which gives the probability of each client being connected to each of the 3 phases.

$$X = ginv(t(P) \times P, tol = 0) \times t(P)) \times B \qquad (16)$$

Where P represents the table with m readings by n customers, corresponding to smart meter readings in each customers' household and B is composed of m readings by 3 phases, corresponding to load totals in each phase per measurement.

In this simulation, the pseudo-inverse with zero tolerance was utilized to compute the matrix inverse, allowing for collinearity and also to allow to run simulations with less readings than number of clients. Also, $t(\,)$ symbolizes the transpose of a given matrix.

Now, the model for PCA is detailed:

$$X = t(-ginv(Cd, tol = 0) \times Ci) \qquad (17)$$

Where, Cd corresponds to the first 3 columns of the U_2 matrix and Ci to all other columns, considering that U_2 is the table corresponding to the last 3 columns of the matrix S given by:

$$S = svd(Z \times t(Z)) \qquad (18)$$

Where Z corresponds to a table with n customers plus 3 by m readings and svd computes the singular value decomposition. It should be noted that in order to compute the inverse, in this model the pseudo-inverse was also applied.

Usually, when comparing algorithms, two ways to evaluate performance are frequently utilized. The first one and theoretically most important is algorithm accuracy. Secondly, processing speed may also have a relevant importance when working with big data such that a slow execution may even compromise real word application of such algorithms.

Algorithm accuracy in the context of this work basically answers the question of how good each algorithm is at correctly inferring customer phase connectivity and is calculated by computing the subsequent formula.

$$Accuracy = \frac{Number\ of\ correct\ guessed\ phases}{Total\ number\ of\ Clients} \qquad (19)$$

Moreover, in order to present more consistent results, Monte-Carlo simulations were conducted with varying numbers of runs in the $[20,50]$ interval. Considering several simulations, the algorithm accuracy is finally considered the average of all runs, given by:

$$Average\ Accuracy = \frac{\Sigma\ Accuracy\ per\ run}{Total\ number\ of\ runs} \qquad (20)$$

In the next section, where results will be presented and analyzed, when the term "accuracy" is referred to actually it means the average accuracy for the simulated Monte-Carlo runs.

IV. Simulation Results

In this section, we summarize the main results of the performed experiments.

A. Noiseless

The first simulation compares the perfomance of both MLR and PCA at infering phase connectivity in an ideal situation where there is no noise added to the problem and thus the totals per phase and per point in time match exactly with the sum of all smart meter readings for that time period.

Average algorithm accuracy results are displayed in Fig. 3.

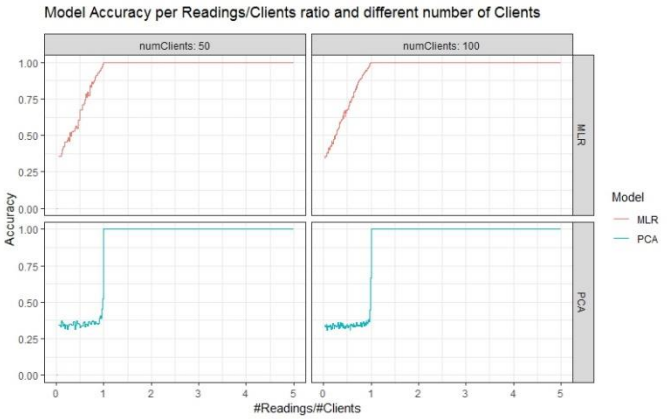

Fig. 3. Average model accuracy for 20 runs without adding errors.

Evidently, both algorithms achieve 100% accuracy as soon as the number of readings per number of clients' ratio is unitary. On the other hand, we observe significant differences between 0 and 1 ratio where MLR's accuracy increases linearly with increasing number of readings while PCA is still random. Note that 33.33% accuracy corresponds to the probability of correctly guessing the phase at random since there are 3 phases.

B. Meter accuracy class

Next, the typical meter accuracy error is included, considering 99.5% accuracy class meters. Although 0.5% meter accuracy error is rather small, it has a slight impact in total model accuracy as can be observed in Fig. 4. In order to achieve approximately 100% meter accuracy, instead of having a number of readings per number of client's ratio of 1, we now need around 1.7 ratio.

Still, both algorithms show roughly the same progression as in the noiseless case.

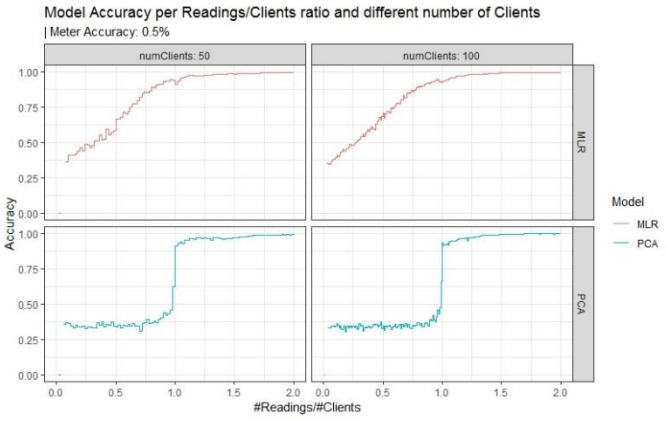

Fig. 4. Average model accuracy for 20 runs with 0.5% smart meter accuracy error.

In order to determine the algorithms' sensitivity to increasing meter accuracy error, results are now presented in the model sensitivity chart displayed in Fig. 5.

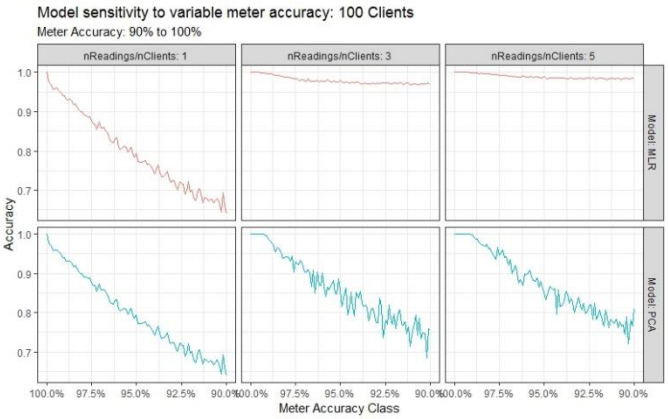

Fig. 5. Model sensitivity to variable meter accuracy for 50 runs and 100 clients.

One can perceive that MLR's accuracy significantly improves when increasing the number of readings from 100 to 300 while PCA seems to show approximately the same linear downwards trend regardless. In fact, given 3 times the number of readings versus clients, MLR never drops below 96% accuracy whereas PCA's accuracy progressively declines until reaching 70% for a 90% meter accuracy class.

C. Clock asynchronism

Introducing the first of the clock errors, results are presented for smart meters with a maximum clock asynchronism of 45 seconds.

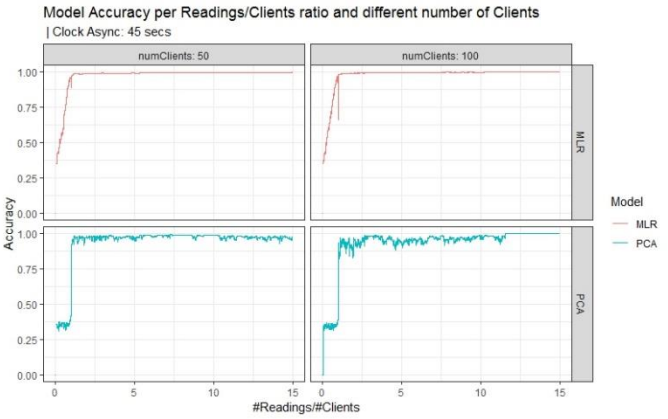

Fig. 6. Average model accuracy for 20 runs with maximum 45 seconds clock asynchronism.

It is clear from the results that MLR suffers little from the simulation of a typical clock asynchronism error. Alternatively, PCA starts to deteriorate its performance, only achieving 100% accuracy when the number of readings is more than 12 times the number of clients, given 100 customers.

To confirm the previous results, each algorithms' sensitivity to clock asynchronism error was then computed and results shown in Fig. 7. In fact, results were in accordance with the previous experiment. Moreover, the outcome is similar to the previous sensitivity analysis on meter accuracy error. MLR's accuracy improves when the number of readings increases but PCA's behavior keeps declining when error increases, although more erratically.

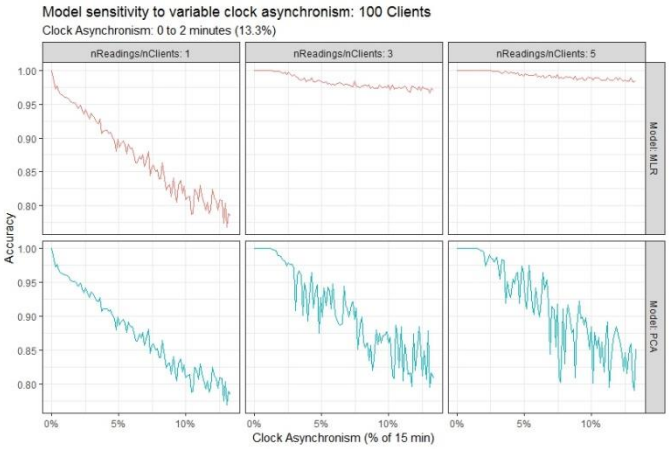

Fig. 7. Model sensitivity to variable clock asynchronism for 20 runs and 100 clients.

D. Clock skew

The second of clock errors is now presented. Considering clock skew errors, Fig. 8 illustrates the results of applying a maximum of 5% skew error.

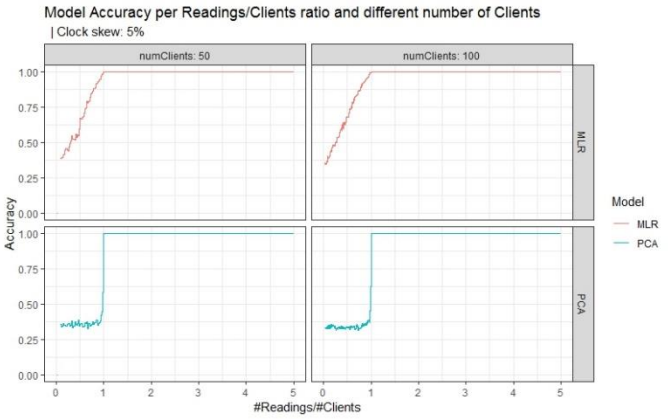

Fig. 8. Average model accuracy for 20 runs with 5% clock skew error.

This result highlights that both models achieve 100% accuracy when the number of readings surpasses the number of clients, even including 5% clock skew error. These findings support the notion that MLR and PCA phase identification models are not influenced by clock skew errors.

To confirm this assumption, each model's sensitivity was plotted in Fig. 9.

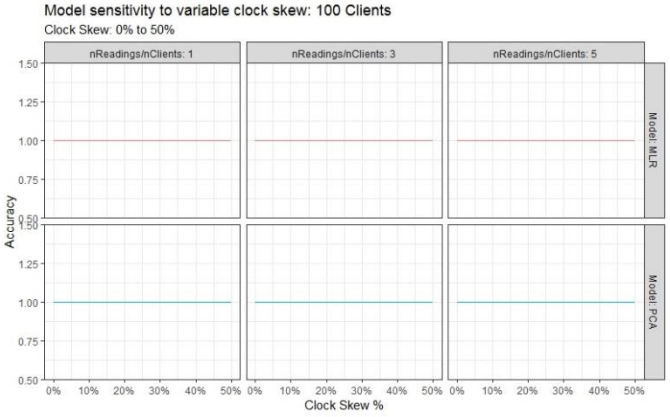

Fig. 9. Model sensitivity to variable clock skew error for 100 clients.

The results revealed tie well with the aforementioned proposition. This appears to be a case of the error having no impact on the correlation between household readings and total load measured at substations because it is constant over time for each smart meter.

E. Copper losses

The following step in the methodology is adding technical copper losses to substation totals to infer the influence of this factor in each models accuracy. Results are provided in Fig. 10.

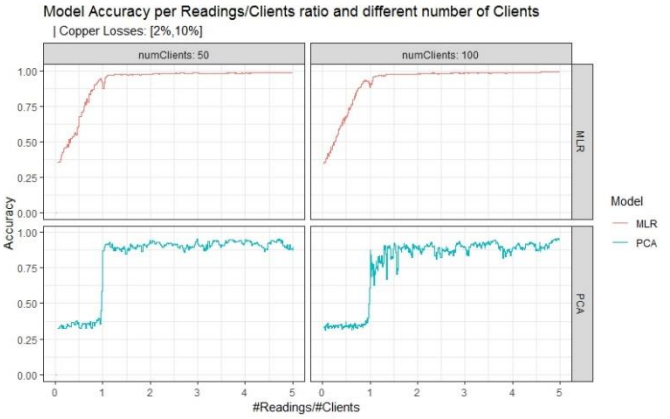

Fig. 10. Average model accuracy for 20 runs with 2% to 10% copper losses, varying quadratically with load.

Following the addition of copper losses, MLR algorithm shows improving results with increasing number of readings, reaching nearly 100% accuracy as the ratio approaches 5. Then again, PCA's performance shows a significant negative impact, averaging around 90% accuracy.

Each algorithm's sensitivity to increasing copper losses is presented in Fig. 11.

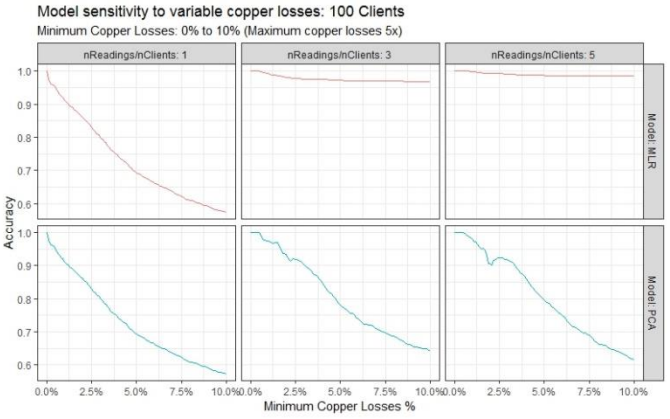

Fig. 11. Model sensitivity to variable copper losses for 20 runs and 100 clients.

Once again, while adding noise to PCA dramatically affects its performance, MLR shows only a minor loss of under 5% in algorithm accuracy when the number of readings per number of clients increases to more than three.

F. Copper losses

The final section on isolated errors presents the results for testing the data with five missing clients.

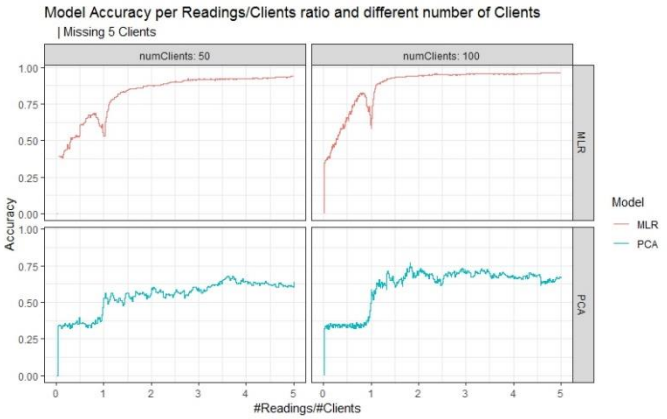

Fig. 12. Average model accuracy for 20 runs with 5 missing clients.

As previously discussed, removing information on clients that only contribute to substation totals and are not fed to the algorithms as smart meter readings has a significant impact on both algorithms performance. However, as has been the case, MLR recovers to nearly 100% accuracy when the number of readings increases to 500 whereas PCA suffers significantly, hovering around 66% accuracy.

Fig. 13 illustrates each model's sensitivity to an increasing number of missing clients, showing that for each customer scraped from the input data algorithm accuracy shows a visible drop. Nevertheless, keeping consistent with results, MLR is much less volatile, even though it drops for the first time below 90% accuracy.

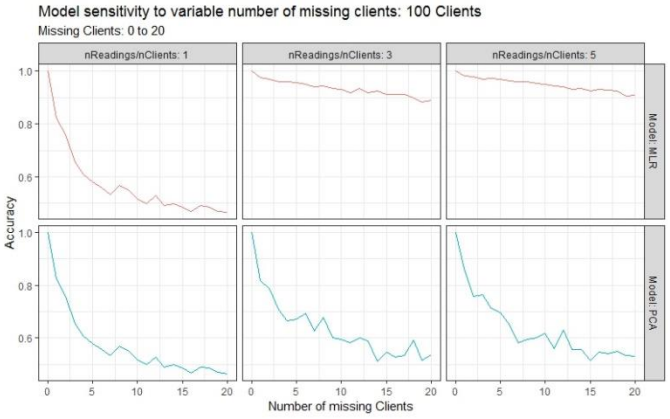

Fig. 13. Model sensitivity to variable number of missing clients for 20 runs and 100 clients.

G. *All errors simultaneously*

Finally, in order to test the robustness of each algorithm under laboratorial conditions simulating real world conditions as closely as possible, both algorithms were tested under all types of error simultaneously.

The next figure illustrates total load per phase, given 3 different plots: 1) Client's total per phase without any errors in red, 2) Total errors included in the simulation in green and 3) the substation phase totals fed to the algorithms in blue, corresponding to the sum of client readings plus errors.

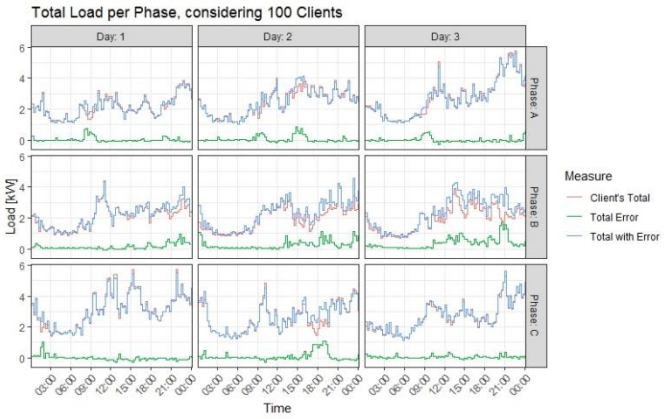

Fig. 14. Total load per phase considering all types of errors.

It is possible to observe from Fig. 14 that total errors are clearly visible even when only typical values, close to reality, are applied.

The ensuing Fig. 15 plots each model's accuracy when all typical errors are included.

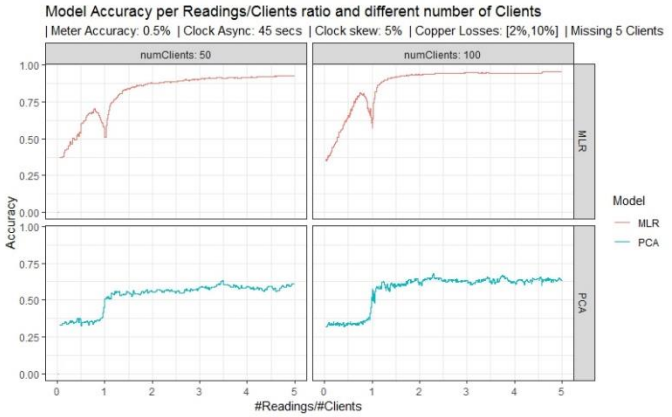

Fig. 15. Average model accuracy for 50 runs with all types of errors.

Excitingly, considering the cumulative effect of all noises included, MLR's performance at inferring phase connectivity shows stellar results. With laboratorial conditions as close as possible to real world data, and maybe some scenarios even more demanding, MLR shows promising results, achieving 98% accuracy with just 5 times the number of readings per number of clients' ratio.

On the other hand, PCA's accuracy hovers close to 60%, and thus, with this simple execution, appears limited for real world implementation.

Interestingly, an inflection in MLR's accuracy when the number of readings nears the number of clients has become evident. Although this effect has been noted in most simulations before, in this example its influence is unavoidable. A possible explanation for this behavior may be that as the number of variables nears the number of available equations, the model is increasingly restricted and thus cannot compute the optimal solution. Another possible explanation for this is the application of pseudo-inverse with zero tolerance to compute the algorithm. Nonetheless, if this algorithm is to be implemented in a real world scenario, further research should be done to investigate the root cause for this inconsistency and possibly deliver a solution.

V. Conclusion

In this dissertation, a new method which applies Multivariate Linear Regression for estimating the customers' phase connectivity was presented, analyzed and its performance compared with a state-of-the-art alternative methods that use Principal Component Analysis techniques. Utilizing real-world data provided by EDP Distribuição for smart meters for a specific location and computing per-phase aggregated phase totals under laboratorial conditions, both algorithms' implementations discarded the need for introducing relaxations or for preprocessing the raw data.

For experimentations without introducing noise, both algorithms always achieve 100% accuracy when the number of readings is greater than or equal to the number of smart meters. However, since in the real-world losses and errors are unavoidable, Monte-Carlo simulations were run with substation data built to replicate typical grid losses, random noise, energy theft, clock skew and clock synchronization errors.

When simulating near-world conditions, Multivariate Linear Regression model successively presented a better performance, consistently achieving 100% accuracy when testing the different types of errors both independently and simultaneously. On the other hand, Principal Component Analysis suffered particularly from energy theft and copper losses, lowering its accuracy to close to 60% when all errors were considered simultaneously.

In order to further assess the robustness of MLR, a simulation with very high error values was performed and, extraordinarily, it still manages to output over 90% accuracy which further increases the confidence in this algorithm for inferring customer phase connectivity in the presence of different kinds of noises.

In addition to delivering better results, MLR's implementation simplicity is a significant advantage in the business context. Moreover, given the fact that the phase identification algorithms presented have a low time complexity, with each simulation in the order of tens of milliseconds, it means a transfer to practice can be attained.

For future works, it would be important to characterize the real business implementation scenario, in order to identify the average number of readings and the average number of clients that are available and, with that information, assess the expected model accuracy. Ideally, given real world secondary substation readings and its connected customers smart meter data, MLR's performance may be assessed without the need to develop error scenarios.

Additionally, it would be relevant to test and compare both models, in similar conditions as tested in this dissertation, but after preprocessing the raw data. If the expected increase of accuracy is significant enough, an increase of implementation complexity in real business applications could be justified.

Finally, in order to perfect MLR algorithm's efficiency, further research should be led to investigate the drop in accuracy when the number of readings to number of client's ratio is unitary.

References

[1] J. P. Satya, N. Bhatt, R. Pasumarthy, A. Rajeswaran, "Identifying Topology of Power Distribution Networks Based on Smart Meter Data," *IEEE Trans. Smart Grid*, pp. 1–8, 2016.

[2] V. Arya, T. S. Jayram, S. Pal, S. Kalyanaraman, "Inferring connectivity model from meter measurements in distribution networks," *Proc. fourth Int. Conf. Futur. Energy Syst.* - e-Energy '13, p. 173, 2013.

[3] G. Sarraf, M. C. A, T. Flaherty, S. Jennings, C. Dann, "2017 Power and Utilities Industry Trends," *PricewaterhouseCoopers*, 2017. [Online]. Available: https://www.strategyand.pwc.com/trend/2017-power-and-utilities-industry-trends

[4] A. Denman, A. Leroi, H. Shen, "How Utilities Can Make the Most of Distributed Energy Resources," *Bain and Company*, Inc., 2017. [Online]. Available: https://www.bain.com/insights/how-utilities-can-make-the-most-of-distributed-energy-resources/

[5] Y. Liao, Y. Weng, M. Wu, R. Rajagopal, "Distribution grid topology reconstruction: An information theoretic approach," *Proc. North Am. Power Symp.* NAPS 2015, 2015.

[6] W. Wang, N. Yu, B. Foggo, J. Davis, "Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data." *Proc. 15th IEEE Int. Conf. Machine Learning and Applications* (ICMLA), Dec. 2016.

[7] B. K. Aakriti Gupta, "Utility AMI Analytics at the Grid Edge: Strategies, Markets and Forecasts," *Mackenzie* 2016. [Online]. Available: https://www.greentechmedia.com/research/report/utility-ami-analytics-at-the-grid-edge#gs.6wGxn5A

[8] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, K. Loparo, "Big data analytics in power distribution systems," *IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf*, June 2015.

[9] W. Wang, N. Yu, Z. Lu, "Advanced Metering Infrastructure Data Driven Phase Identification in Smart Grid," *Proc. GREEN*, Sept. 2017.

[10] D. K. Chembe, "Reduction of Power Losses Using Phase Load Balancing Method in Power Networks," *Proc. World Congr. Eng. Comput. Sci.*,Vol. I, 2009.

[11] C. Lueken, P. M. S. Carvalho, J. Apt, "Distribution grid reconfiguration reduces power losses and helps integrate renewables," *Energy Policy*, vol. 48, pp. 260–273, 2012.

[12] S. J. Pappu, N. Bhatt, R. Pasumarthy, A. Rajeswaran, "Identifying Topology of Low Voltage (LV) Distribution Networks Based on Smart Meter Data," *IEEE Trans. Smart Grid*, vol. 3053, no. c, pp. 1–1, 2017.

[13] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. John Wiley & Sons, Inc, 2002.

[14] S. Narasimhan, "Deconstructing Principal Component Analysis Using a Data Reconciliation Perspective," *Computers & Chemical Engineering*, vol. 77, no. 9 pp. 74-84, June 2015.

[15] "RStudio Version 1.0.143." 2016.

[16] V. J. Hodge, J. Austin, "A Survey of Outlier Detection Methodologies," *Artif. Intell. Rev.*, vol. 22, pp. 85–126, 2004.

[17] Satec (Australia) P. Ltd, "Electricity metering accuracy explained," 2014. [Online]. Available: https://www.ecdonline.com.au/ content/ electrical-distribution/article/electricity-metering-accuracy-explained-372339275#axzz5TwvC4ABT

[18] Microsemi, *The New Role of Precise Timing in the Smart Grid*, White Paper Revision 1.0, 2017. [Online]. Available: https:// www.microsemi .com/document-portal/doc_view/133267-the-new-role-of-precise-timing-in-the-smart-grid

[19] V. J. Hodge, J. Austin, "A Survey of Outlier Detection Methodologies," *Artif. Intell. Rev.*, vol. 22, pp. 85–126, 2004.

[20] Satec (Australia) P. Ltd, "Electricity metering accuracy explained," 2014. [Online]. Available: https://www.ecdonline.com.au/ content/ electrical-distribution/article/electricity-metering-accuracy-explained-372339275#axzz5TwvC4ABT

[21] Microsemi, *The New Role of Precise Timing in the Smart Grid*, White Paper Revision 1.0, 2017. [Online]. Available: https:// www.microsemi .com/document-portal/doc_view/133267-the-new-role-of-precise-timing-in-the-smart-grid

[22] ECI Telecom Ltd., *Fighting Electricity Theft with Advanced Metering Infrastructure*, March 2011

Biographies

João L. T. Santos received the electrical and computer engineering degree from Instituto Superior Técnico (IST), University of Lisbon, Portugal, in 2012. He is currently finishing his M.Sc. degree in the same area and same university. Since 2013, he has been with Deloitte Consulting SA., at the Strategy and Operations Department for products, services, utilities, and resources industries.