Phase Identification in Low Voltage Grids: Experimenting over different data analytics approaches under laboratorial conditions

João Luís Testa Santos

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Pedro Manuel Santos de Carvalho

Examination Committee

Chairperson: Prof. Rui Manuel Gameiro de Castro

Supervisor: Prof. Pedro Manuel Santos de Carvalho

Members of the Committee: Prof. José Manuel Dias Ferreira de Jesus

November 2018

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

The conclusion of this master thesis represents a significant milestone in my academic and professional lives, allowing me a greater freedom of mind and spirit, releasing me from hidden bindings and becoming the stepping stone to reaching my full potential. This accomplishment would not have been possible without the help and support of so many people, to whom I present my deepest gratitude and acknowledgment.

Firstly, to Professor Pedro Carvalho for entrusting me with the responsibility to see this work through, in uncommon conditions. For his permanent availability and guidance along the process, while allowing for my independent research. Finally, for his wisdom and knowledge, not only in statistical and mathematical theory but also in broad practical experience.

Also, I would like to thank my current employer Deloitte Consultores, SA for providing me with tools, experiences and work ethics that allowed me to face this challenge in a timely manner. Furthermore, I appreciate conceding me this opportunity to take a leave from work to conclude my academic background, as I recognize the complexity in managing a tight scheduling.

To my great friend and colleague João Machado for giving the kick-off to this project and proving all his support, counseling and companionship in all respects. Also, a special thanks to Alexandre Dias for his fellowship during the course of this endeavor.

To my family, for their permanent support and applying the right amount of pressure and incentive to keep me coming back to achieve this milestone in my education.

Lastly and most importantly, to Andreia, for her patience, support and incentive, her precious help and for always being by my side.

To all those I haven’t mentioned and have helped me fulfill this journey in one way or another.

Resumo

A rede de distribuição de baixa tensão frequentemente não possui informação atualizada sobre a conectividade à fase de cada cliente. Este facto origina obviamente ineficiências na gestão do equilíbrio trifásico, o que por sua vez pode gerar ineficiências operacionais tais como aumento de perdas ou desequilíbrios de tensão desnecessários. Contudo, com a instalação de *smart meters* e a consequente disponibilização de dados de consumos de energia de clientes a intervalos de tempo pré-determinados é possível estimar a ligação à fase de cada cliente, assumindo que está também disponível informação sobre o consumo agregado por fase nas subestações, com a mesma resolução de tempo.

Nesta tese, um conjunto de abordagens tutoriais de *data analytics* que permitem identificar a conectividade à fase subjacente dos clientes com base no seu histórico de consumos e totais agregados por fase nas subestações foi estudado. Com base nesse estudo, um novo método que aplica a regressão linear multivariada foi implementado e o seu desempenho comparado com um método proposto na literatura, que utiliza Análise dos Componentes Principais [1].

A experimentação é realizada não só em (*i*) condições laboratoriais, nas quais a informação agregada por fase nas subestações é construída de forma a replicar perdas da rede típicas, ruído aleatório, roubos de energia da rede e erros de assincronismo e enviesamento do relógio, mas também (*ii*) em dados reais fornecidos pelo incumbente em Portugal, EDP Distribuição – Energia, SA para uma localização específica.

Palavras-chave: Identificação de fase, Redes Inteligentes de Baixa Tensão, Topologia de Rede, Regressão Linear Multivariada, Análise dos Principais Componentes

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.

Key-words: Phase identification, Low Voltage Smart Grids, Smart Meters, Network Topology, Multivariate Linear Regression, Principal Component Analysis

Table of Contents

List of Tables

List of Figures

Glossary

MLR: Multivariate Linear Regression

PCA: Principal Component Analysis

DER: Distributed Energy Resources

AMI: Automated Metering Infrastructure

CM: Connectivity Model

PIS: Phase Identification System

GPS: Global Positioning System

PMU: Phasor Measurement Unit

SVM: Support-Vector Machines

DBSCAN: Density-Based Spatial Clustering Applications with Noise

MIP: Mixed Integer Programming

OLS: Ordinary Least Squares

SVD: Singular Value Decomposition

1. Introduction

1.1. Motivation

Phase identification is a critical input to the grander problem of phase load balancing. 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 manpower 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 and thus this thesis seeks to present a novel and simpler method for phase identification, utilizing Multivariate Linear Regression (MLR), and compare its performance to the state-of-the-art method proposed in [13] which utilizes Principal Component Analysis (PCA).

1.2. Objective

The objective of this work is to present a set of tutorial approaches to identify underlying customer phase-connectivity from:

1) Their time series of energy measurements and;
2) Their aggregated per-phase energy measurements.

This smart meter data is gathered at a pre-defined time-resolution, set by the power utility companies, usually between 15 min and 1 hour.

In this paper a method is proposed which applies Multivariate Linear Regression to infer phase connectivity and its performance compared with literature provided Principal Component Analysis [1].

In order to achieve a simpler implementation, both methods are developed without need for relaxations or pre-processing. Though the application of such techniques may improve the performance of both algorithms, accurate and significant results are still obtained while keeping the implementation simple, intuitive and easily replicable in power utility companies.

Method evaluation will be established by measuring model accuracy under different data sources and errors. Data sources include both laboratorial conditions where aggregated per-phase energy measurements data is built from daily consumer profile samples and real-world data provided by Portugal’s incumbent EDP Distribuição – Energia, SA, for a specific location. Different errors will be added to increase test robustness, namely typical grid losses, smart meter accuracy, energy theft, and clock skew and synchronization errors.

1.3. Document Structure

Firstly, the current section concludes Chapter 1 which is an introduction into the topic discussed in this thesis.

Secondly, Chapter 2 introduces the background and historical solutions to tackle the phase identification problem as well as exhibit state-of-the-art methodologies proposed to solve this problem.

Chapter 3 delves into the mathematical problems considered and implemented in this work, beginning by detailing each of the models and then presenting a qualitative comparison of the proposed methodologies, identifying theoretical pros and cons of each one.

Afterwards, in Chapter 4, the implementation methodology is discussed, starting by explaining input data utilized in the subsequent experimentations, then describing how different noises and losses were modulated on top of this input data, and finally characterizing the testing framework utilized for gauging the effectiveness of the different algorithms and describing the calculation method for key measures.

In Chapter 5, the performance of the implemented methodologies under different scenarios is presented and analyzed, according to the designed testing framework.

Finally, Chapter 6 concludes this work with key observations and suggestions for future work related to this paper.

2. Background and State-of-the-Art

2.1. Background

Historically, a number of different approaches have been used to solve the phase identification problem. These approaches may be classified between hardware-based or software-based.

Introducing the hardware-based methodologies, the simplest solution is to deploy smart meters with Power Line Communication capabilities which can be used to directly communicate with secondary substations comparing both PLC and substation measurements to correctly identify the measured phased. This solution, however, is not always viable either because many smart grid networks have already been deployed without such capabilities or because the incremental cost of such equipment may increase total investment to be too high for investors [14].

Previous to smart grids, Phase Identification Systems (PIS), for instance the one introduced in [15], relied on Global Positioning System (GPS) to compare phase measurements between a known reference phase location and mobile units used in unknown field locations. Despite having been in service since 2004, the costs and manpower required by these solutions presents a significant downside. Furthermore, many of such methods necessitate customer availability to provide access to specific meters inside households [14].

Other methods, such as the one presented by Chen *et al*. [16] which refers a method that incorporates a microprocessor and usage of GPS to infer phase connectivity and the one discussed in [17] which applies signal injection devices suffer from the same drawbacks.

With the advent of smart grids using AMI and Phasor Measurement Units (PMU), software-based methods, often referred to as big data analytics or machine learning, have become increasingly popular and are now considered state-of-the-art techniques to infer phase connectivity. Located at households in a smart grid, or at important nodal points, these devices measure different data at regular time-intervals and synchronized with each other [9].

As the number of smart meter installations is forecasted to surpass 1.1 billion worldwide by 2022, large volumes of diverse data will feed data analytical approaches to phase identification, improving its performance and accuracy. Data most commonly available includes electricity consumption, measured every 15 minutes to 1 hour, voltage magnitude, geographical information, asset health monitoring or, with micro-PMUs, time-synchronized measurements with phase angles [9].

2.2. State-of-the-Art

State-of-the-Art methodologies to infer phase connectivity and network topology through data analytics differ not only on the type of input data exploited, time-series of voltage magnitude or time-series of energy measurements, but also on the algorithms and machine learning techniques proposed to solve the problem.

In the literature, the majority of researchers have proposed considering the time-series of voltage magnitude correlation between customers and feeders on the same phase. The intuition is that the impedance and loads on the distribution network are inherently unbalanced and therefore results in unbalanced line currents and voltages. The implication is that customers with the same phase will present trajectories of voltage time-series with similar behavior to each other [7].

In machine learning, techniques may be categorized as supervised, semi-supervised and unsupervised. Supervised techniques require an accurate subset of customer to phase connectivity data with which to train the algorithms to properly identify the correct phase for the unknown customers. Semi-supervised learning also requires some labeling data to train the algorithm, however the training to testing dataset ratio is much smaller. On the other hand, unsupervised learning techniques do not rely on training the algorithm with correct data.

In [18], the performance of different classification machine learning methods, support-vector machines (SVM), label propagation and clustering by k-means, one from each technique respectively, is compared, by training the algorithms on an existing and partially correct phase connectivity model, with voltage magnitude measurements. While SVM presents the highest accuracy, it requires accurate phase connectivity for a subset of customers, which may not always be available. Oppositely, the k-means clustering unsupervised technique requires no prior knowledge of phase connectivity and may be more robust to noisy labels.

In fact, diverse clustering techniques have been proposed. In [7], a two-step clustering algorithm is proposed. First, a linear dimensionality reduction technique, PCA, is used to extract key feature vectors from the raw time-series of voltage measurements. Afterwards, the k-means clustering algorithm is applied to identify customers belonging to the same phase. In contrast, [10] uses a nonlinear dimensionality reduction technique, t-SNE, to feed the unsupervised clustering algorithm density-based spatial clustering applications with noise (DBSCAN). A comparison of both algorithms is presented in [19], where GIS data is also leveraged to improve results. Despite presenting accurate results and not relying on correct labels for training, clustering methodologies suffer from needing a small scale field validation to attribute a single phase to each cluster.

Other methodologies also consider different approaches to deduce network topology from voltage correlation. In [6], the topology of a distribution network is constructed via an information theory based algorithm, Chow Liu algorithm, from voltage measurements. In [14], the maximum spanning tree graph theory is used, by applying the Prim algorithm to find the maximum spanning tree of the complete graph. A hybrid solution between the Prim and Chow Liu algorithms is developed in [20]. Non-synchronous voltage data is considered in [21] to find the related maximum likelihood by estimating the inverse covariance matrix while incorporating prior information on line statuses via a maximum a-posteriori approach. In [22], harmonic voltage correlation is proposed to determine the phase connectivity of customers, based on a correlation analysis with the Fischer Z transform.

In [23], a two-step algorithm is developed by estimating the concentration matrix using a garrote-type estimator, and then deriving that same matrix to infer the most probable electrical network. This method however only works reliably without noise and with a reasonably small phase error.

Subsequently are described two methods have been proposed that use both voltage and energy measurements. In [24], linear regression based algorithms are applied to find basic voltage drop relationships, and in [25] network topology identification is achieved through estimation of the network voltage sensitivity matrix and posterior application of graph theory (Prüfer Sequence). Information on both voltage and energy time-series is however not always available.

Finally, methodologies using time-series of energy measurements in kilowatt-hour are presented. These approaches are based upon the principle of energy conservation which implies that total energy supplied by a feeder in each phase must be equal to the energy consumed by all the households connected to that phase plus errors. One disadvantage of using this principle is that results are more sensitive to unmetered loads when compared to voltage-based algorithms.

In [3], mathematical optimization is used by proposing different relaxations to Mixed Integer Programming (MIP) formulations. Conversely, this implementation is reported to be computationally intensive. In [1], the authors propose applying PCA and its graph-theoretic interpretation to infer phase connectivity from the time-series of energy measurements, comparing its performance to [3].

In this thesis, work is focused on the time-series of energy measurements since not only are solutions utilizing this data scarcer, but also because it was the data made available to the development of this project. The technique proposed in this effort which is based upon Multivariate Linear Regression is compared with Principal Component Analysis presented in [1], since it is reported to be the most recent and best performant solution to the phase identification problem while using energy measurements data.

3. Predictive Models in this Study

3.1. 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 [26].

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 $\mathrm{x}_\mathrm{i}^\mathrm{T}\beta$ is the inner product of vectors x_i and β.

Specifically, MLR is the implementation that best fits the problem discussed in this thesis. 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 $(q + 1) \times p$.

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. $cov\left(y_i, y_j\right) = 0$, 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:

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

3.2. 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:

$$CZ = 0 \tag{6}$$

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:

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

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

$$S_R(U_2^T) \sim S_R(C) \tag{8}$$

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

$$U_2^T Z = 0 \tag{9}$$

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:

$$U_2^T Z = \hat{C} Z = Q \hat{C} Z = 0 \tag{10}$$

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:

$$Z = \begin{bmatrix} Z_d \\ Z_i \end{bmatrix} \tag{11}$$

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:

$$\hat{C} = \begin{bmatrix} \hat{C}_d \\ \hat{C}_i \end{bmatrix} \tag{12}$$

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

$$\hat{C}_d Z_d + \hat{C}_i Z_i = 0 \tag{13}$$

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

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

Where $\hat{R}$ is the $(n_d \times n_i)$–dimensional regression matrix, proven to be unique in [27].

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

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

3.3. 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:

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

Since in most of the simulations $m > n$, $O(n^2m)$ 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. However, preliminary results presented in chapter 4.5 do not follow the expected result since our implementation of the MLR algorithm doesn’t apply the Cholesky factorization by virtue of simplicity.

4. Methodology

4.1. Overview

In order to accomplish the scope of this thesis, the algorithms in analysis were implemented in R Studio [28] 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 simple flow describing the program implementation is represented in the Figure 1 below:

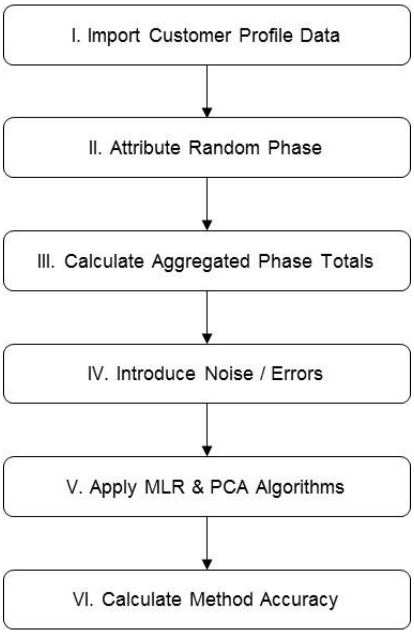
I. Import Customer Profile Data
II. Attribute Random Phase
III. Calculate Aggregated Phase Totals
IV. Introduce Noise / Errors
V. Apply MLR & PCA Algorithms
VI. Calculate Method Accuracy

Figure 1 - Phase identification program workflow

4.2. Input Data

For the development of this thesis, 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 EDP Distribuição, SA. 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 power measurements is in kW. Some data cleansing was necessary as some profiles had missing readings for some hours. Where information was unavailable, it was considered zero. However, daily profiles which had no data or it was always null were removed as they were irrelevant for the problem at hand.

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. Figure 2 represents the load diagrams for a sample of 2 customers, spawning over 3 consecutive days.

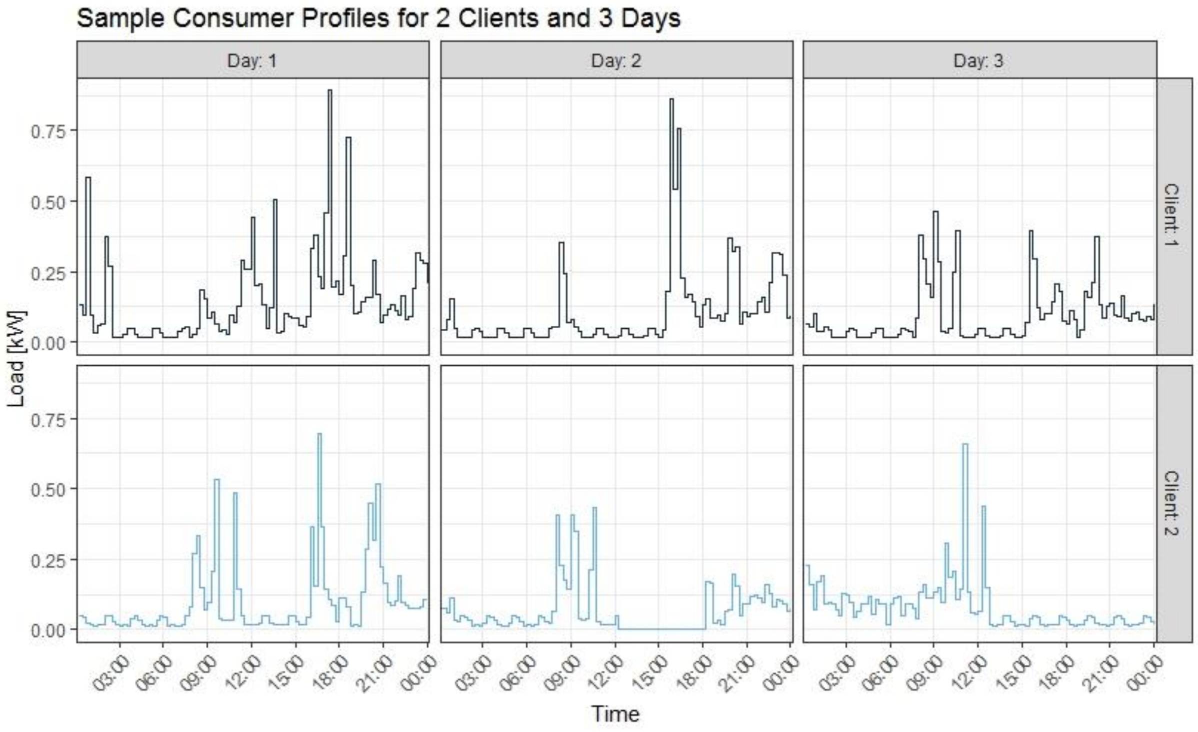

Figure 2 - Sample consumer profiles for 2 clients and 3 days

In the above figure, 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 [29].

The next step is to randomly allocate a phase to each customer, following a uniform distribution. Figure 3 illustrates the total number of customers per phase, considering a test run with 100 clients.

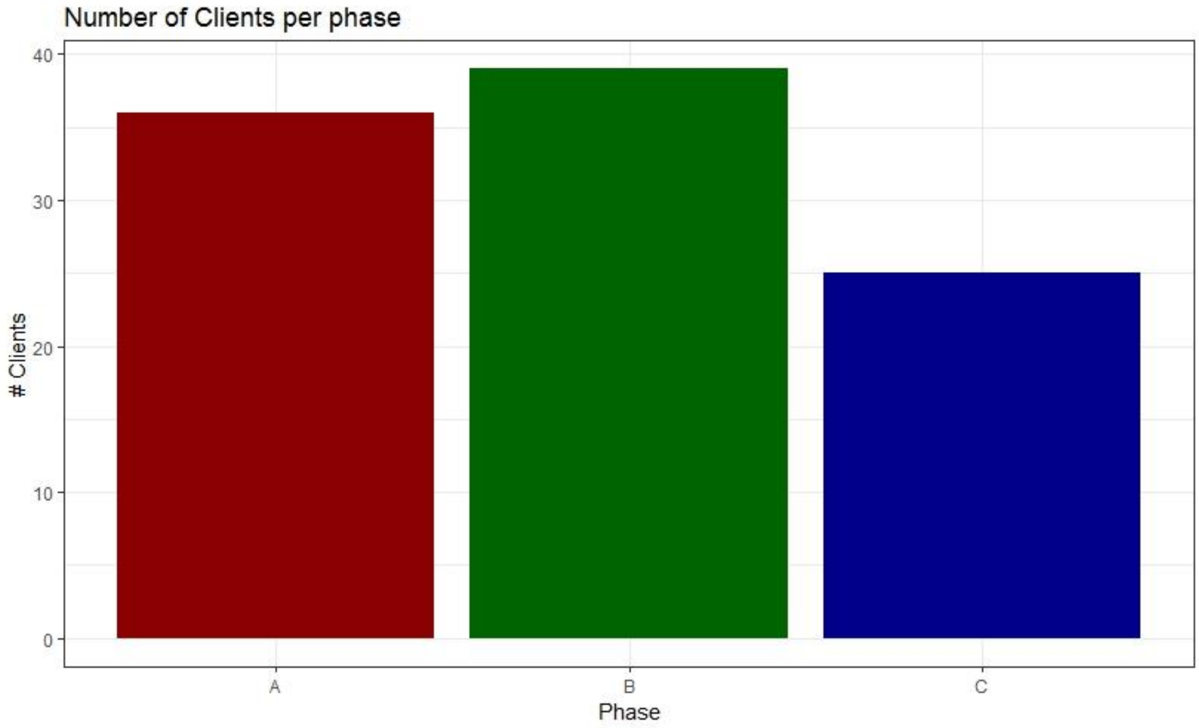

Figure 3 - Number of clients per phase

It is possible to observe that, despite the fact that a uniform distribution was used, there is, in this example, a noteworthy unbalance in the number of clients per phase. In fact, simulations were run to investigate the impact of phase unbalance in the proposed algorithms with no registered influence.

Subsequently, load profiles were aggregated according to their new allocated phase. Figure 4 simulates the readings from a secondary substation, following an aggregation considering Figure 3's allocation.

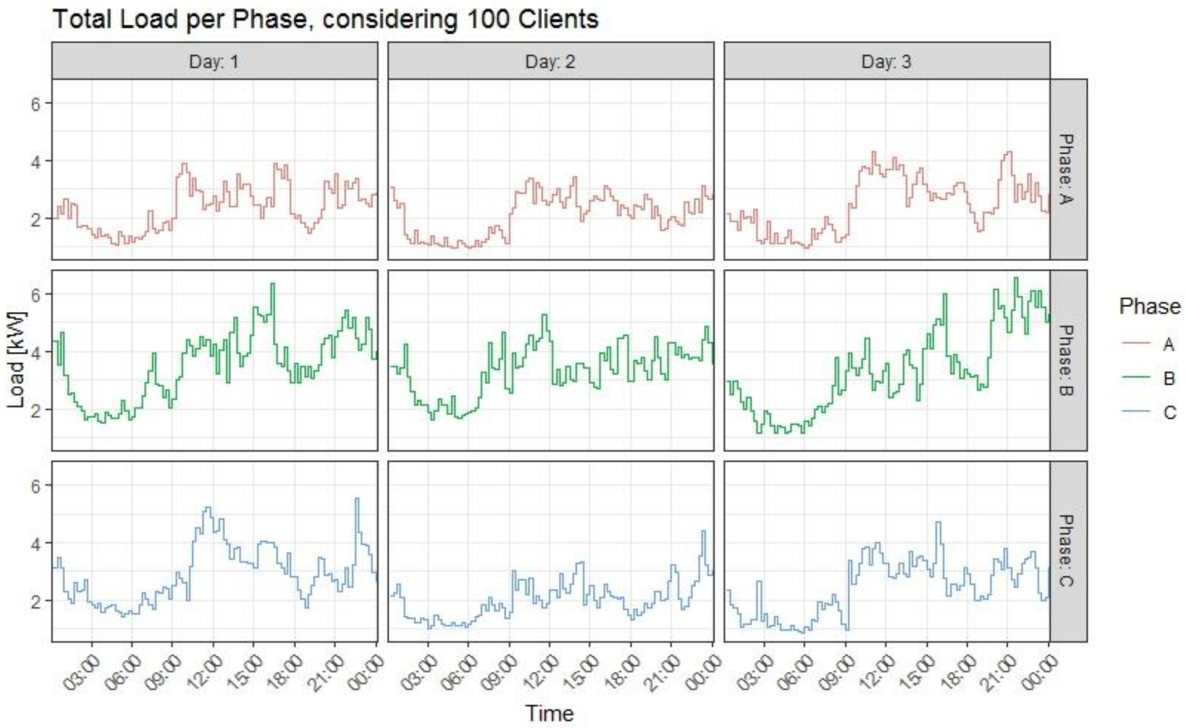

Figure 4 - Total load per phase, considering 100 clients

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, such errors are unavoidable in real situations and thus the following section explores different types of errors to be considered in the analysis.

4.3. Noise modeling

In this chapter, different types of errors will be introduced. A brief explanation of each type of noise will be presented and a typical value introduced. In this work, five kinds of noise were considered:

- **Meter accuracy class**: meters are regulated to have at least 99.5% accuracy
- **Clock asynchronism**: instead of clocking the load at the same time, different meters may have a slight walk
- **Clock skew**: may result when a clock’s frequency differs from the true clock
- **Copper losses**: due to resistive capacitance which results in heating, etc.
- **Missing Clients**: due to theft, missing data, etc.

For each of the errors stated above, detail will be presented as to how they are calculated and algorithm accuracy will be computed isolating each type of noise. The results section will then conclude by aggregating all types of noise simultaneously. Results will be show for a typical error and for a more critical one in order to facilitate the comparison between both algorithms’ accuracy.

4.3.1. 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 [30].

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 modelled 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, as represented in Figure 5.

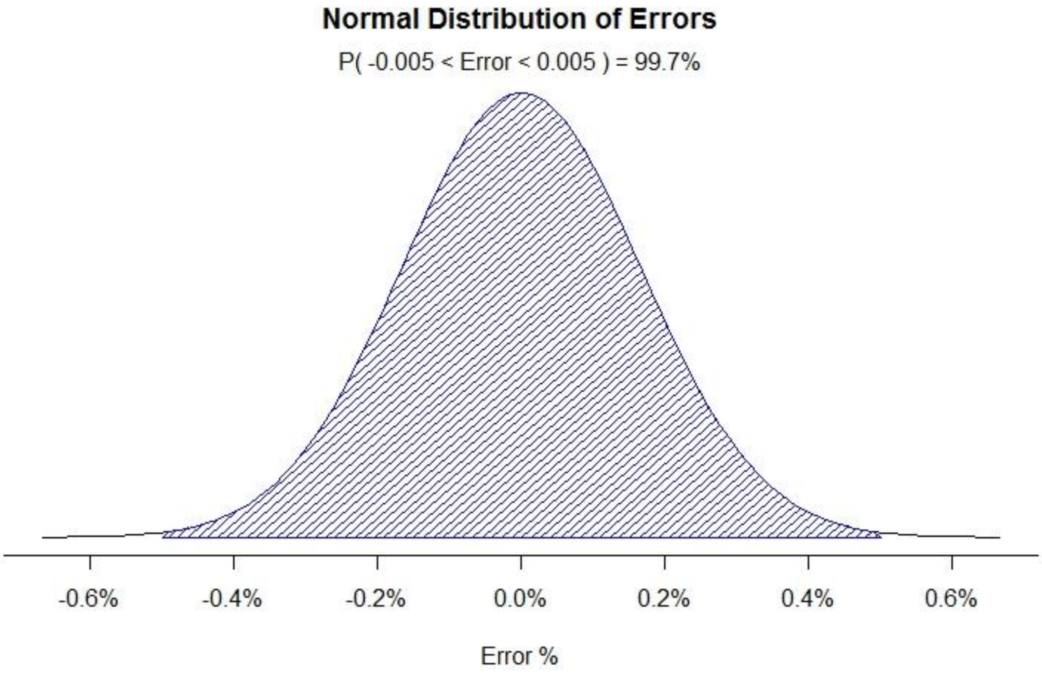

Figure 5 - Normal distribution of meter accuracy error for 0.5 accuracy class

The Figure 6 below shows a sample of 2 clients and 3 days readings including meter accuracy errors randomly selected from the above distribution when a typical meter accuracy error of 0.5% is applied.

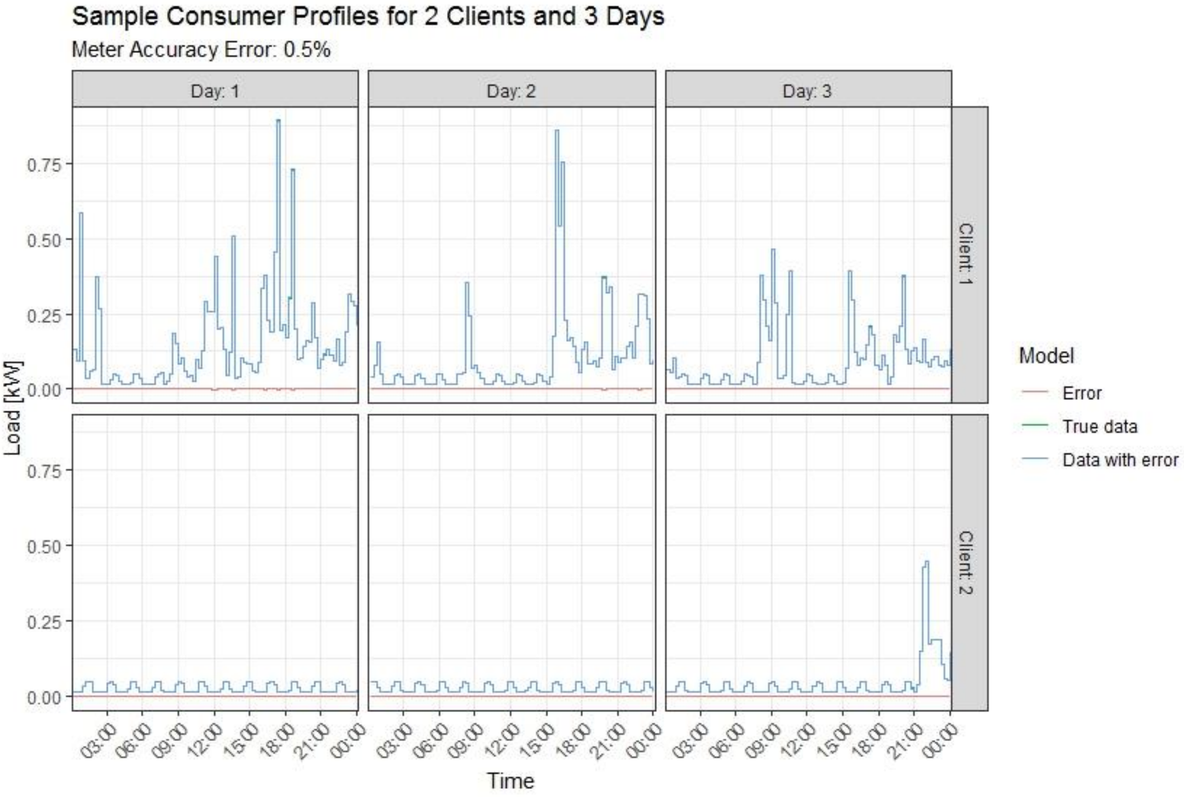

Figure 6 - Sample consumer profiles with meter accuracy error

It is hardly observable any variations on the error value (red line).

4.3.2. 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 [31].

Following V. Arya *et al*. [3] implementation, to simulate clock asynchronism, each meter is made erroneous by adding a random Gaussian walk. Instead of clocking the load after every $\Delta\mathrm{t}$ units, the k^{th} measurement clocks the load for the time interval $[T_{K-1}, T_K]$ where $\mathrm{T}_K = T_{K-1} + N(\mu = \Delta\mathrm{t}, \sigma = \mathrm{f}\Delta\mathrm{t}), \mathrm{f} \in [0, 2.23]\%$. 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%. This will be considered the typical asynchronism error.

Figure 7 shows the normal distribution of clock asynchronism errors considering the conditions stated above. The ensuing Figure 8 displays a sample of 2 clients and 3 days readings including randomly selected asynchronism errors from such a distribution.

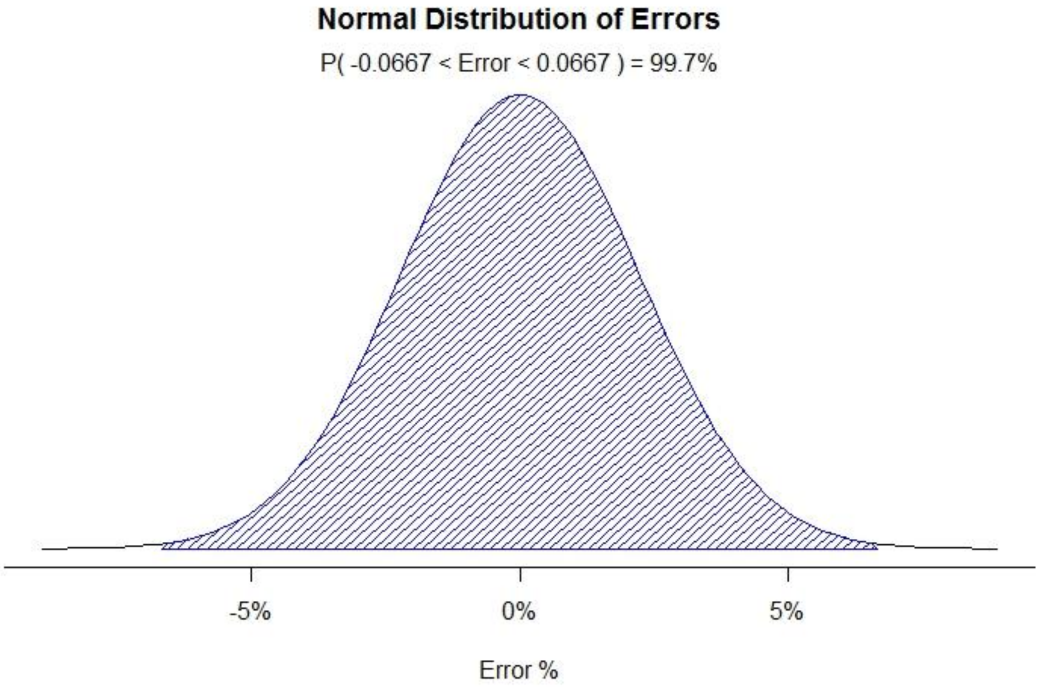

Figure 7 - Normal distribution of clock asynchronism error

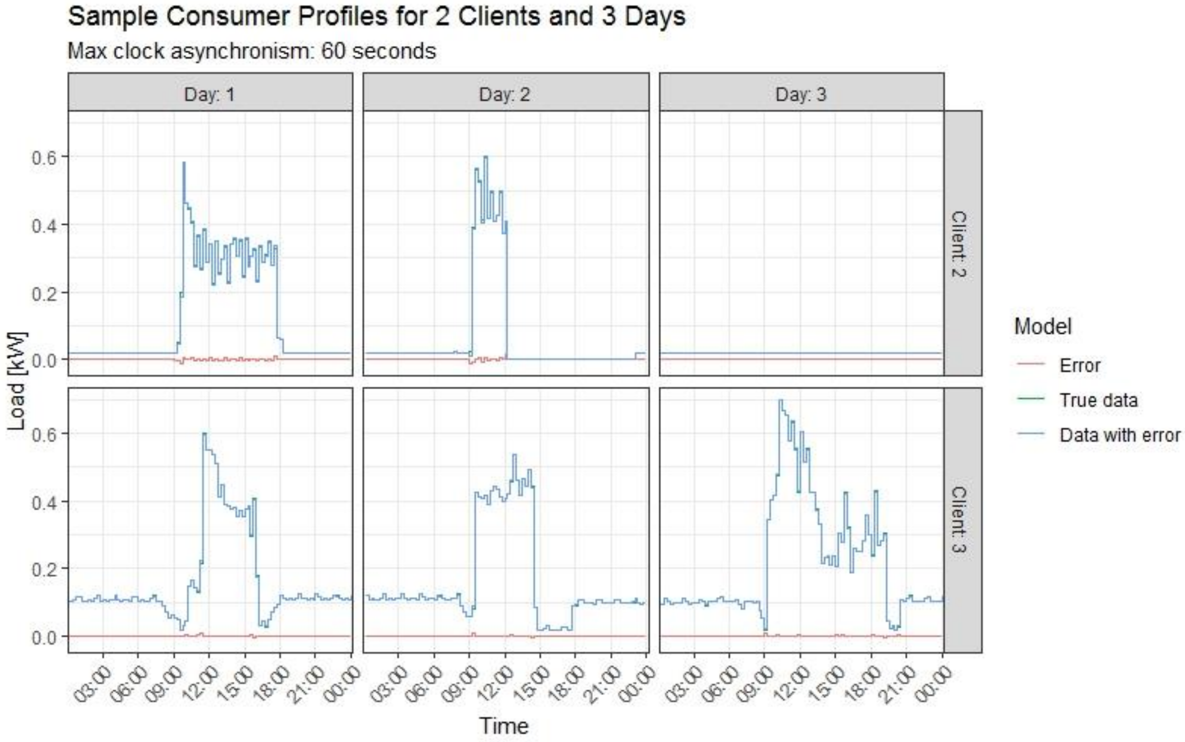

Figure 8 - Sample consumer profiles with clock asynchronism errors

In the figure above, it is possible to observe slight errors, mostly when there are big differences in consecutive readings for true data. This representation is, however, a very high error considering smart meters are usually synchronized from within 1 second of an acceptable time reference [31].

4.3.3. 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, variations in the manufacturing process, power supply variations and different load capacitance [32].

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 $[-\mathrm{f}\Delta\mathrm{t}, \mathrm{f}\Delta\mathrm{t}], \mathrm{f} \in [0, 30]\%$.

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 [3].

Figure 9 below shows the normal distribution from which clock skew errors were randomly selected.

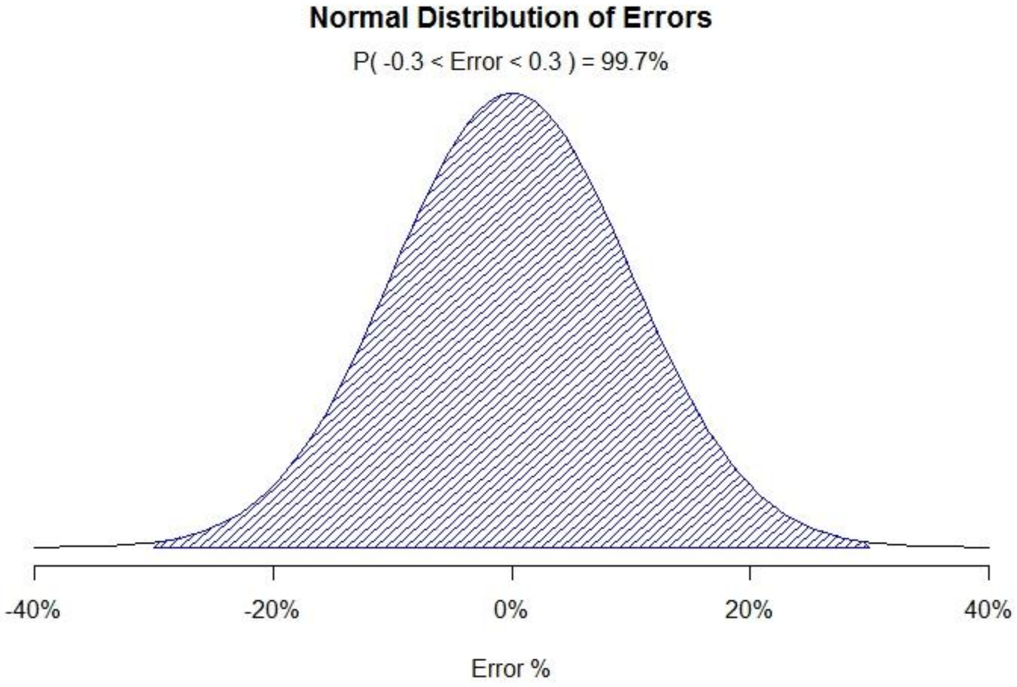

Figure 9 - Normal distribution of clock skew error

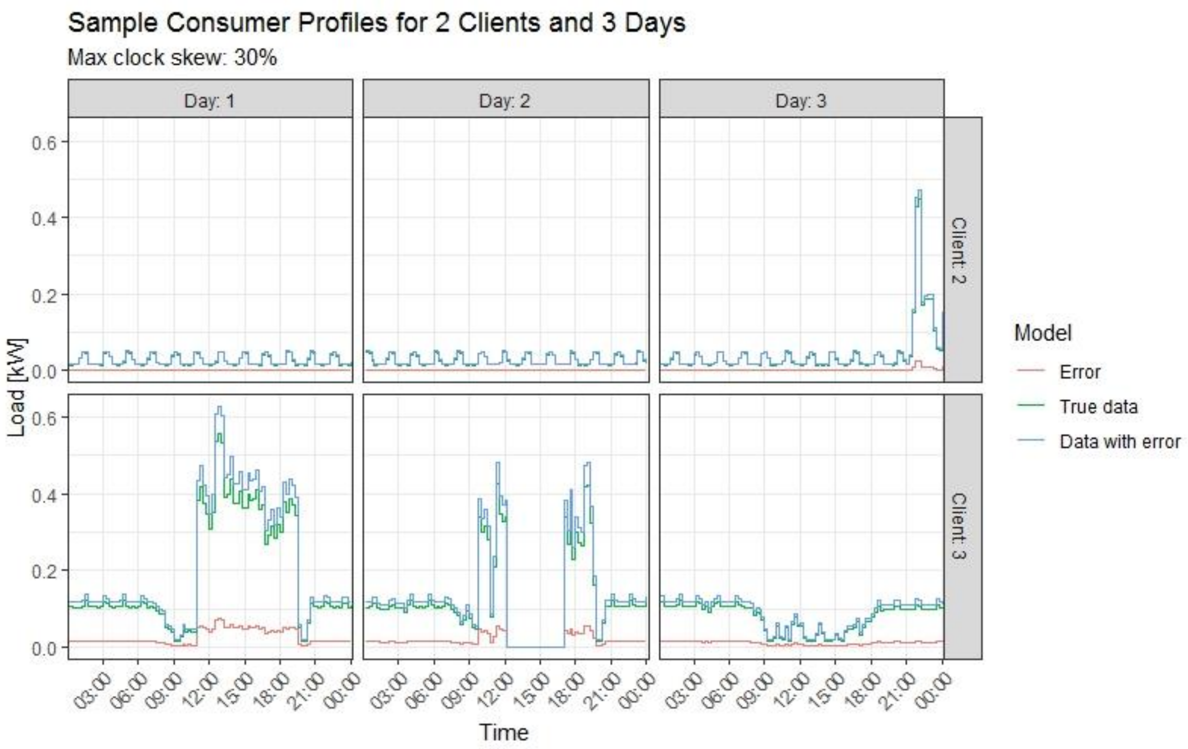

Figure 10 - Sample consumer profiles with clock skew error

From analysing Figure 10 above it is possible to observe that comparatively with the errors displayed before, clock skew error is, in this example, much more visible. However, because the error percentage is constant over time for a single smart meter clock it won't have a significant impact in results as will be discussed in the following chapter.

4.3.4. 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 [33].

In this segment technical copper losses are 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 [34].

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. A sample of 2 customers and 3 days data which is displayed in Figure 11.

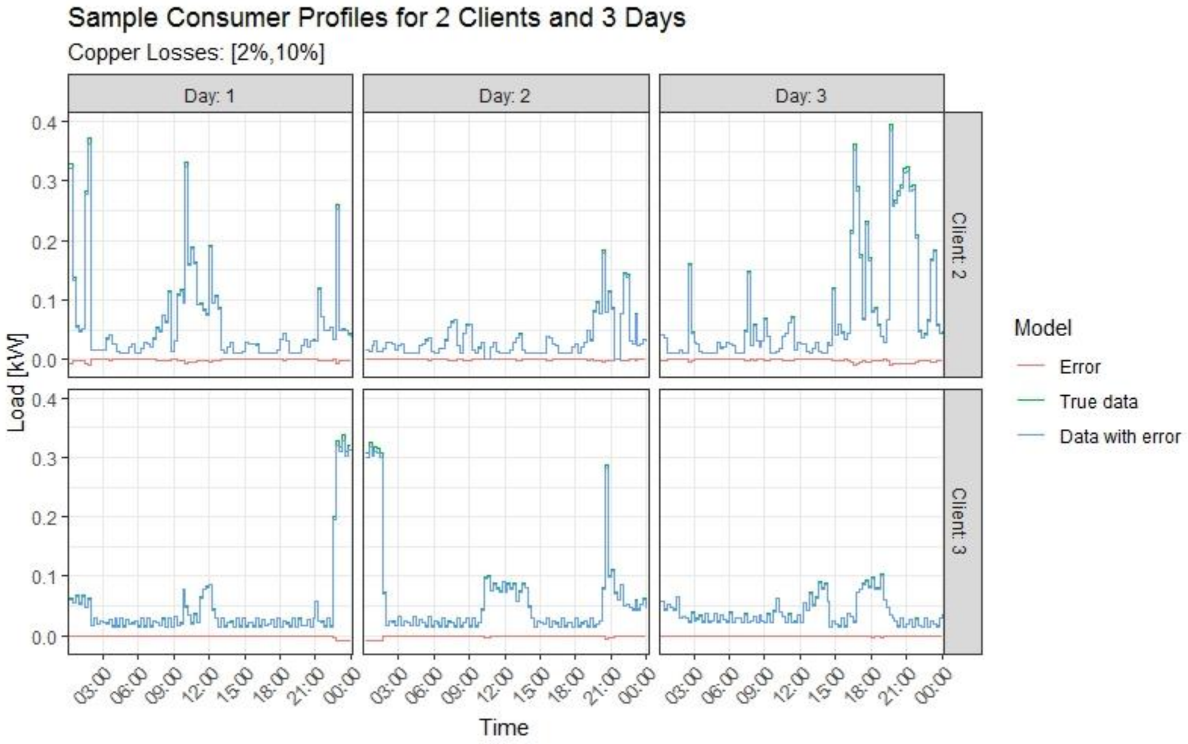

Figure 11 - Sample consumer profiles with copper losses

It is possible to observe that since copper losses vary quadratically with load, errors are more visible when load increases. It should be noted that while all previous errors could be either positive or negative, copper losses are evidently always negative. Figure 12 below shows the totals per phase when considering copper losses for 100 clients.

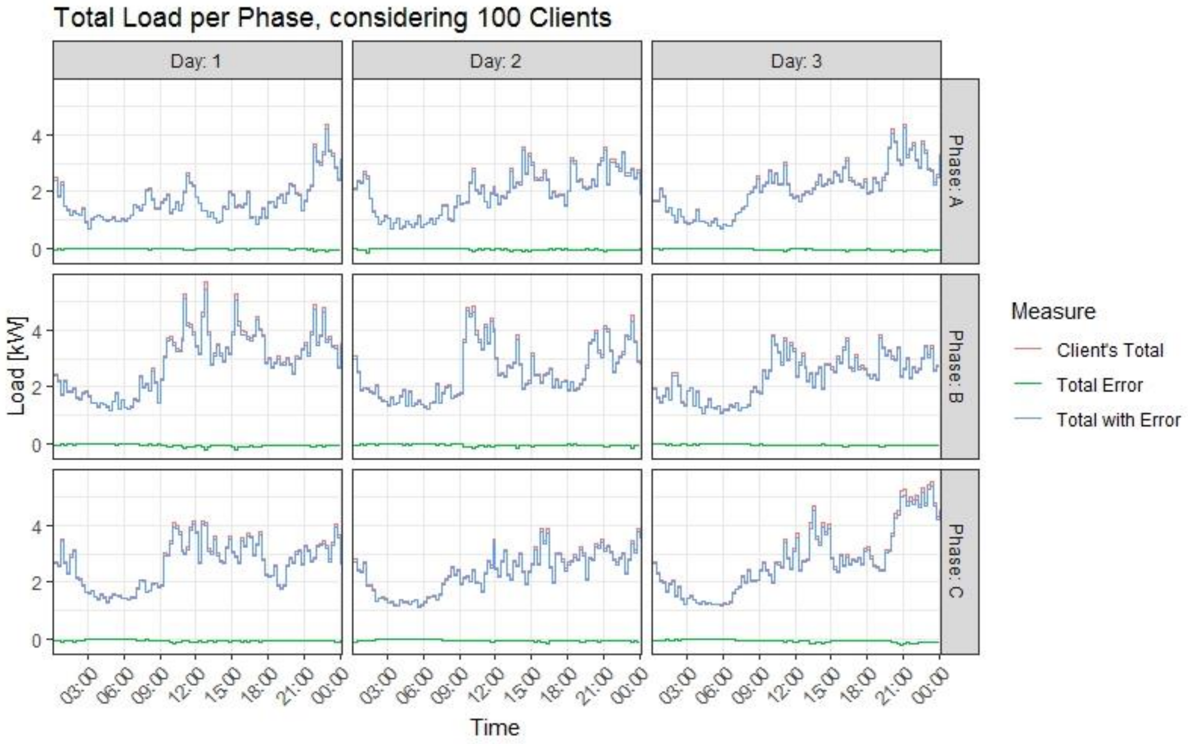

Figure 12 - Total load per phase, including copper losses

4.3.5. 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 unauthorised 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 [35].

While in developed countries secure networks experience only around 1-3% electricity theft, developing countries have been shown to have much higher theft percentages [36].

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%.

Figure 13 shows the phase totals for 100 customers, including 5 missing clients.

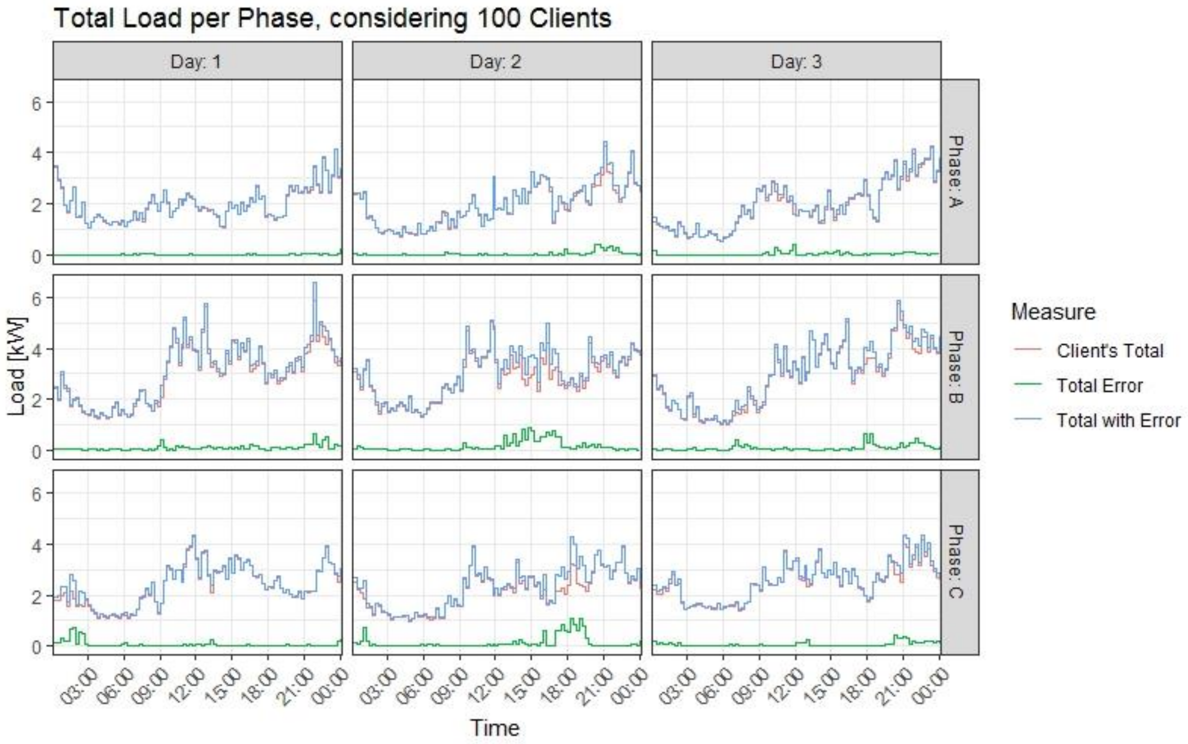

Figure 13 - Totals per phase including 5 missing clients

It is possible to see from this example that electricity theft mostly impacts phase C and phase B. 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.

4.4. Model implementation

Following chapter 3, this chapter explains how the MLR and PCA algorithms are applied to the problem of phase identification and implemented in RStudio.

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

$$X = ginv(t(P) \times P, tol = 0) \times t(P)) \times B \tag{16}$$

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:

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

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:

$$S = svd(Z \times t(Z)) \tag{18}$$

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.

4.5. Performance measures

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.

$$Accuracy = \frac{Number\ of\ correctly\ guessed\ phases}{Total\ number\ of\ Clients} \tag{19}$$

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 to be the average of all runs, given by:

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

In the next chapter where results will be presented and analyzed, when the term “accuracy” is referred to in point of fact it means the average accuracy for the simulated Monte-Carlo runs.

Regarding the algorithms’ time complexity and processing speed, as previously explained in chapter 3.3, because our problem employs few data points its relevance is negligible. Nonetheless, a time performance simulation is presented in Figure 14, given the following input data:

- Number of clients: 150
- Number of readings: [0, 1500]
- Number of runs per data point: 10
- Errors: all 5 errors were added with typical values
- Time is measured in seconds, as the average of all runs

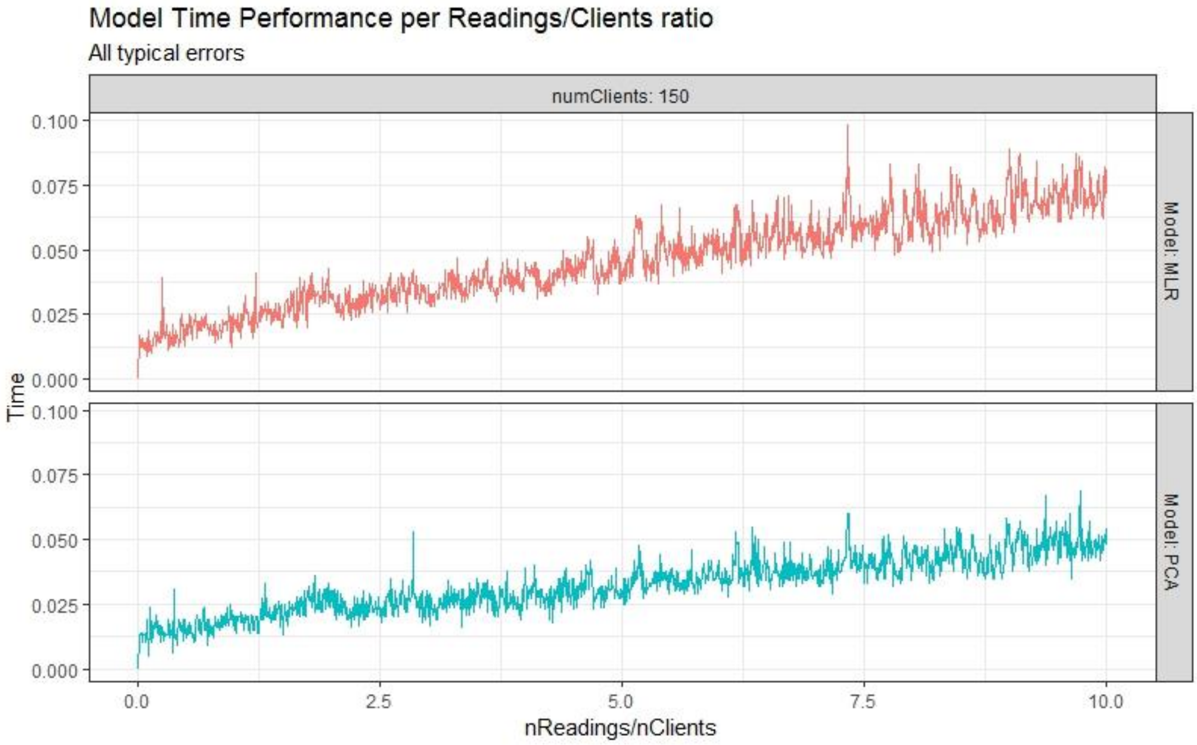

Figure 14 - Time performance with increasing number of readings

Surprisingly, despite the theoretical time complexity given in chapter 3.3, PCA appears to process faster. Taking as an example the last data point, for 150 clients and 1500 readings, PCA averages 0.05 seconds while MLR takes approximately 0.07 seconds.

Appreciatively, both algorithms run in the order of tens of milliseconds which indicates they are viable for real world applications.

4.6. Simulation framework

In order to assist in the navigation of the test results displayed in the next chapter, a simulation framework was developed. It rests mainly on two types of graphics:

I. **Model accuracy** – This graphic summarizes the evolution of each algorithms' accuracy in guessing the correct customer phase for a set of test runs.
    a. **Grid rows** – Each row in the chart grid corresponds to one algorithm: MLR and PCA
    b. **Grid columns** – Each column plots the results for a given number of clients: 50 or 100
    c. **Horizontal axis** – Number of readings per number of clients ratio: the algorithms were fed with unitary increments of number of readings from 1 reading to up 5 times the number of clients
    d. **Vertical axis** – Algorithm accuracy %

An example for visual reference is presented below.

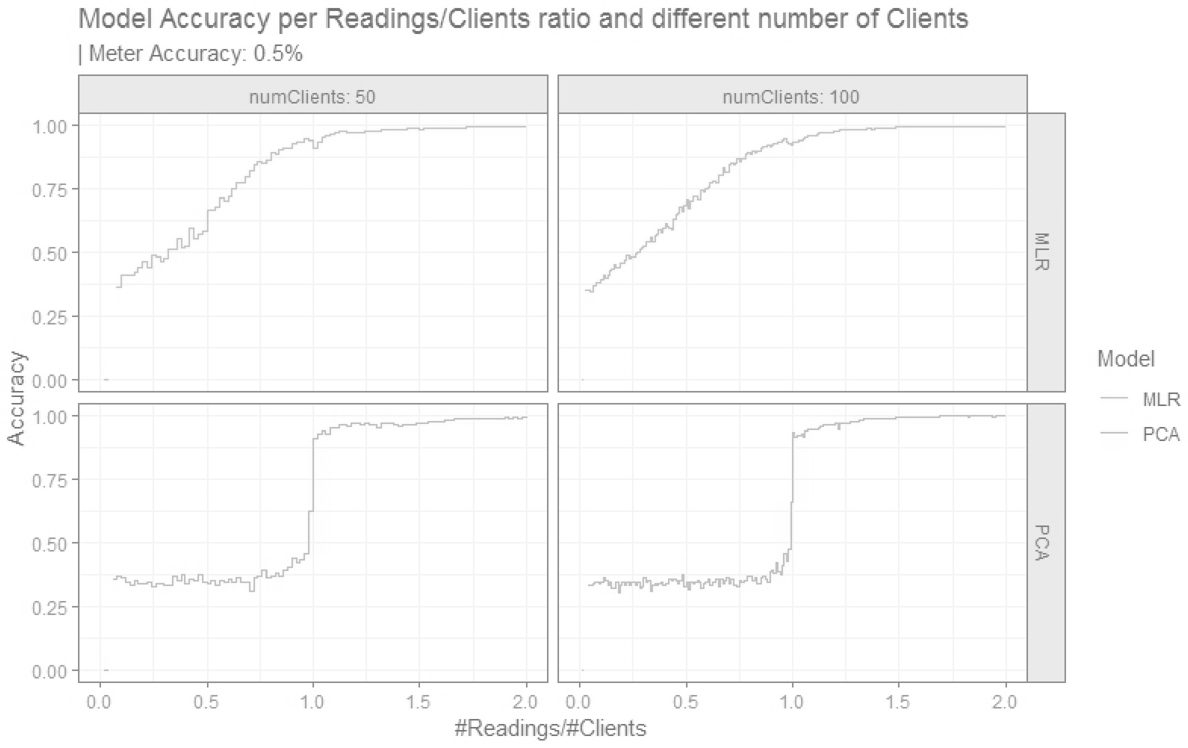

Figure 15 - Example of model accuracy template

II. **Model error sensitivity** – this graphic represents each model's sensitivity to variations in noise, by showing its accuracy with increasing error values. The purpose of this analysis is to find the error value that generates an average of 95% accuracy, henceforth referred to as *critical error*, which will then be applied to the model accuracy chart.
    a. **Grid rows** – Each row in the chart grid corresponds to one algorithm: MLR and PCA

b. **Grid columns** – Each column plots the results for a given number of readings per number of clients ratio: 1, 2 or 3

c. **Horizontal axis** – Increasing error value, dependent on each type of error: axis starts with no error

d. **Vertical axis** – Algorithm accuracy %

e. **Other constants:** for this analysis, data for 100 clients was considered

Another example for visual guidance is provided next:

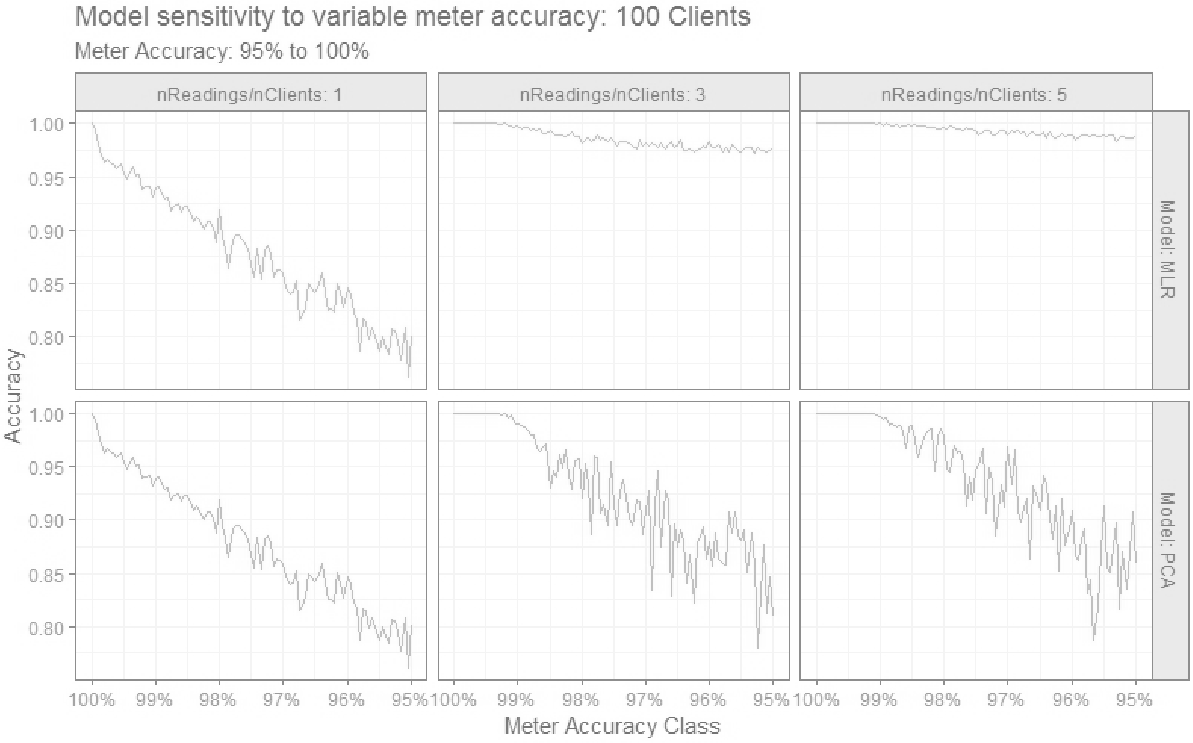

Figure 16 - Example of model sensitivity template

The ensuing Figure 17 presents the simulation framework followed throughout this work, starting from the top, moving clockwise. In can be read as follows:

- **Step 1 – Noiseless -** The algorithm accuracy is presented for both MLR and PCA
- **Steps 2 to 6 – Single error added:** firstly the algorithms' accuracy with a typical error is presented, afterwards the model error sensitivity analysis is run and critical error identified. Finally, the impact of running the model accuracy analysis using the critical error is shown
- **Step 7 – Cumulative errors:** in step 7 the model accuracy results for running all the errors simultaneously are shown, for both typical and critical errors

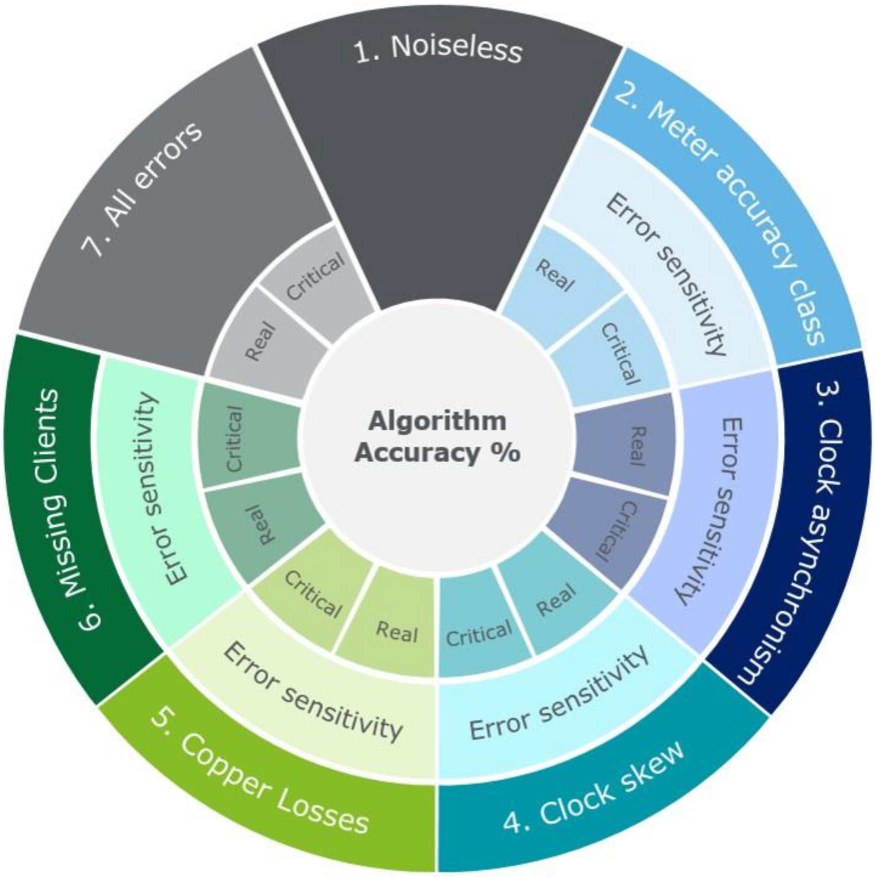

Figure 17 - Simulation framework

5. Results

In this chapter, the results are presented and analyzed according to the simulation framework defined in the previous chapter.

5.1. 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.

The following input variables were applied for this case:

- Number of Clients: 50 and 100
- Number of Readings: [0, 500]
- Number of runs per datapoint: 20

Results are displayed in Figure 18 below.

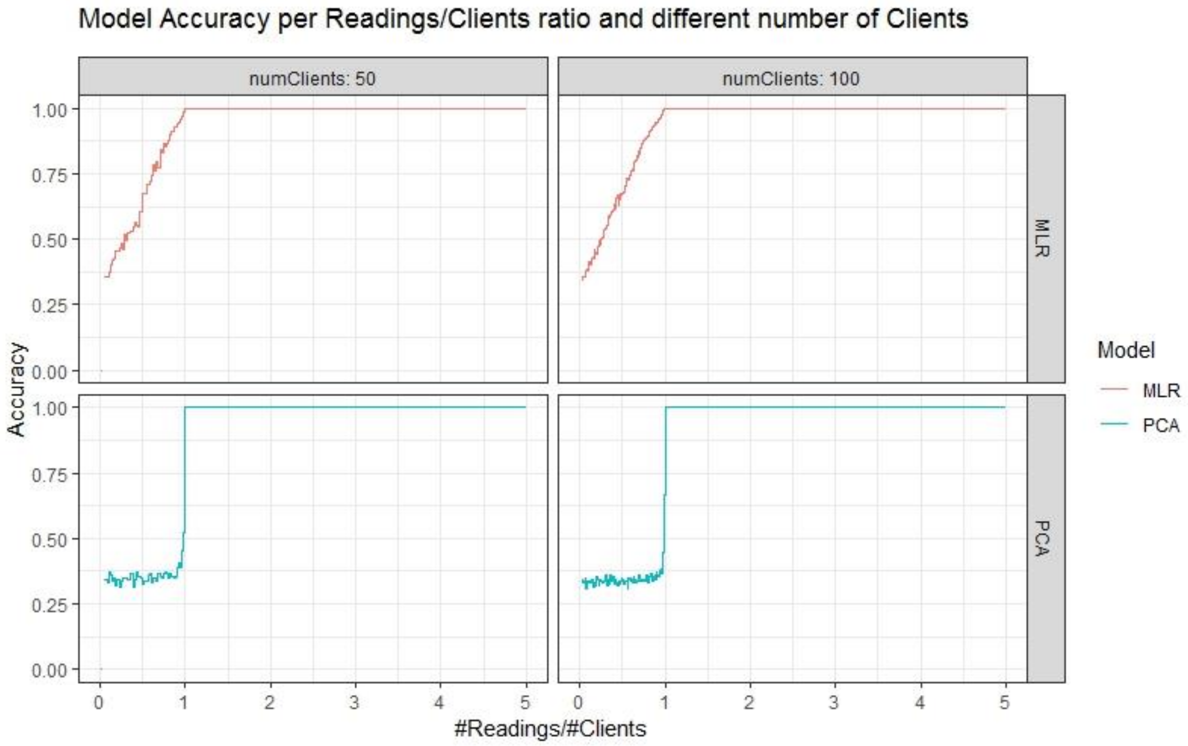

Figure 18 - Model accuracy - Noiseless

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. In real use cases, this difference may not be impactful since achieving a ratio of 1 corresponds to approximately 1 to 2 days of smart meter data if readings are taken every 15 min.

Table 1 presents the necessary minimum number of days collecting data to achieve 100% accuracy for the noiseless case, depending on the number of clients and frequency of readings.

Number of Clients	Minimum number of days (15 min Readings)	Minimum number of days (30 min Readings)
96	1	2
192	2	4
288	3	6
384	4	8

Table 1 - Minimum number of days data to infer phase connectivity without noise

5.2. Meter accuracy class

5.2.1. Typical meter accuracy class

Next, the typical meter accuracy error is included as described in the 4.3. Noise Modelling chapter and results are displayed in Figure 19.

The following input variables were applied for this simulation:

- Number of Clients: 50 and 100
- Number of Readings: [0, 200]
- Number of runs per data point: 20
- Meter accuracy class: 99.5%

Although 0.5% meter accuracy error is rather small, as presented previously in Figure 6, it has a slight impact in total model accuracy. 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.

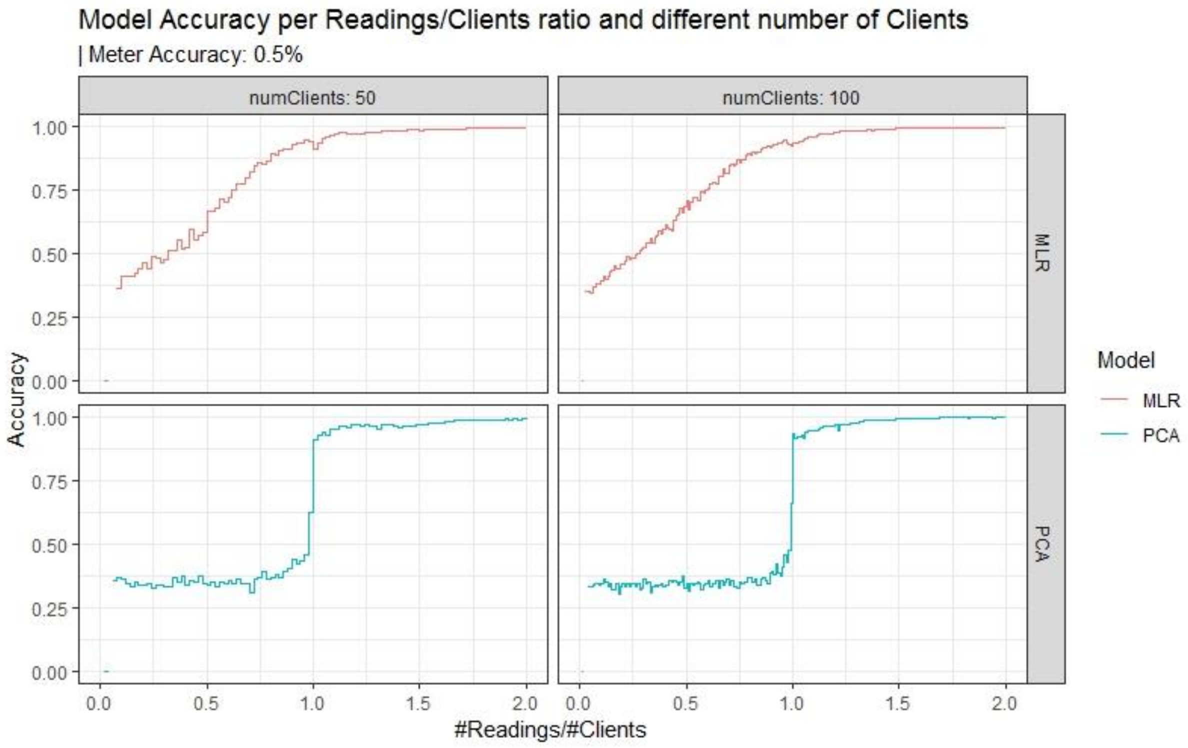

Figure 19 – Model accuracy with 0.5% meter accuracy error

5.2.2. Critical 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 Figure 20.

The plot was computed with the subsequent input data:

- Number of Clients: 100
- Number of Readings: 100, 300 and 500
- Number of runs per data point: 50
- Meter accuracy class: [90%;100%]

It is possible to observe 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.

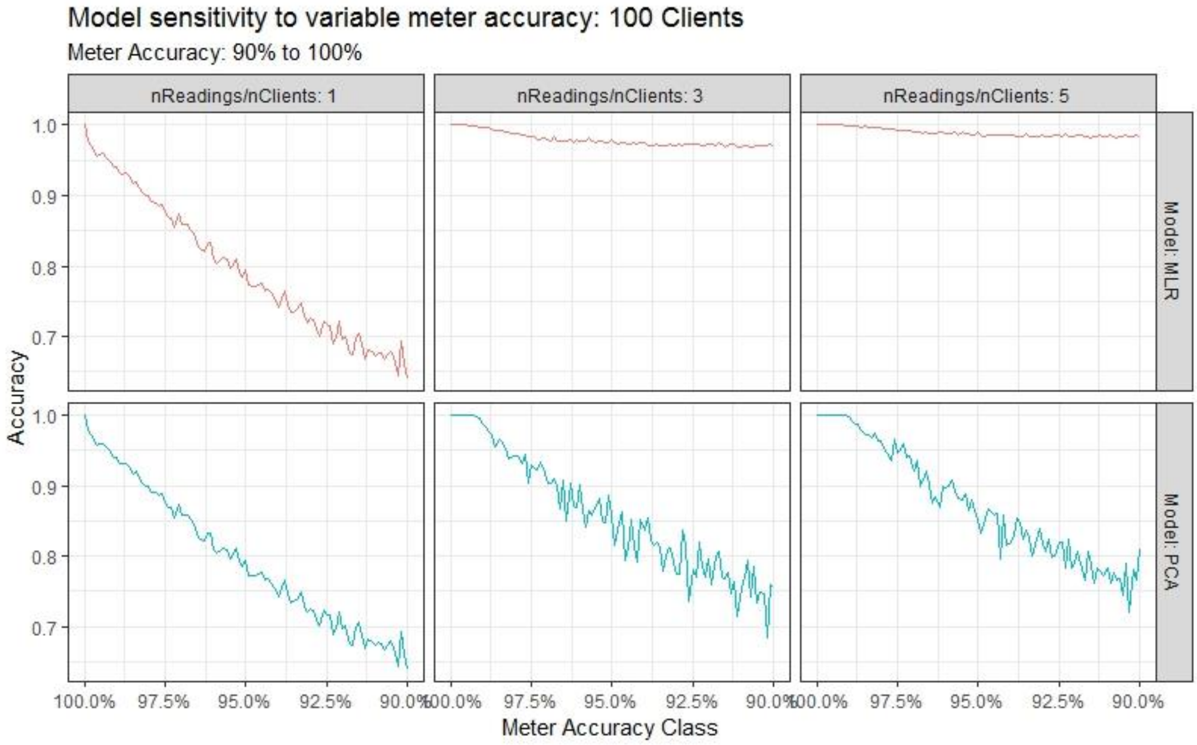

Figure 20 - Model sensitivity to meter accuracy

Given these results, 97.5% meter accuracy was considered to be the critical accuracy error where both algorithms show approximately 90% accuracy when the number of readings equals the number of clients.

The ensuing figure plots each algorithms' accuracy for the critical error with the following variables:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Meter accuracy class: 97.5%

We can see from Figure 21 that MLR shows considerable better accuracy than PCA when the critical error is applied. When the number of readings is twice or more than the number of clients, MLR's accuracy averages 98.7% while PCA's average accuracy is approximately 91.7%.

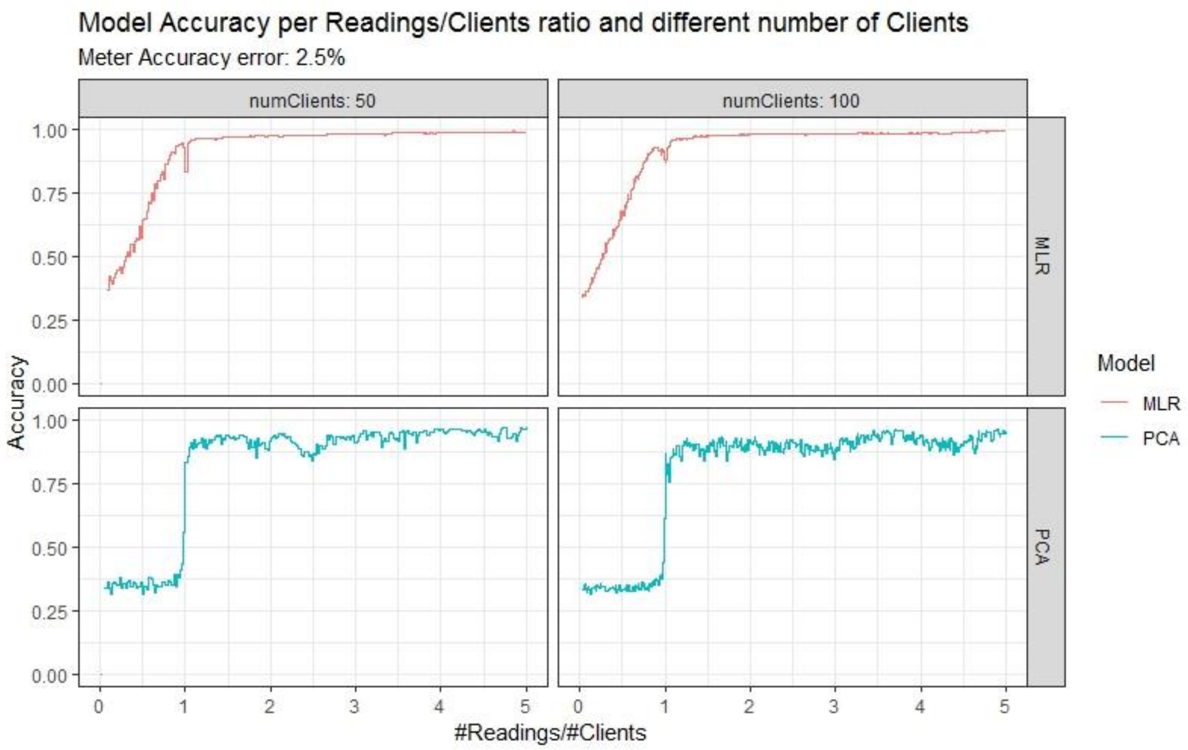

Figure 21 - Model accuracy with meter accuracy error: 2.5%

As a result, it is possible to conclude that, in the presence of a high meter accuracy error, MLR outperforms PCA at inferring phase connectivity.

5.3. Clock asynchronism

5.3.1. Typical clock asynchronism

Introducing the first of the clock errors, results are presented for smart meters with a slight clock asynchronism.In this simulation, variables were configured as follows:

- Number of Clients: 50 and 100
- Number of Readings: [0, 1500]
- Number of runs per datapoint: 20
- Clock asynchronism: [-45, +45] seconds

Results are displayed in Figure 22 below.

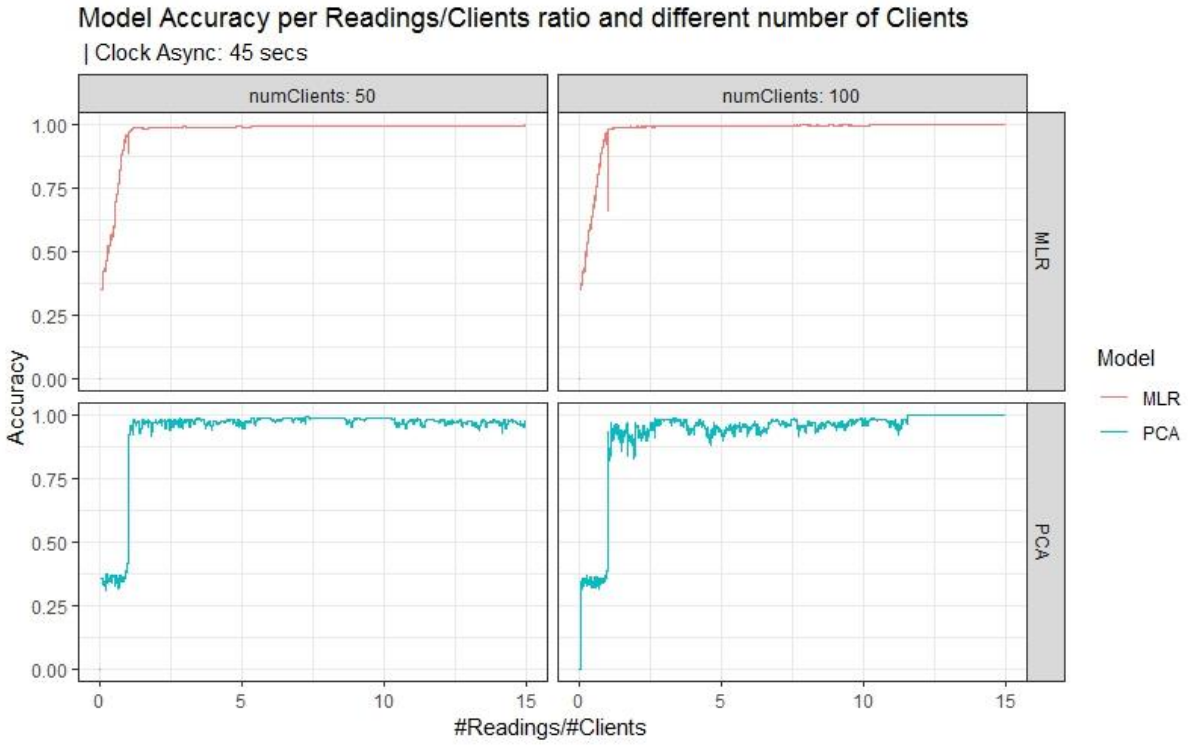

Figure 22 - Model accuracy with max 45 seconds clock asynchronism

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.

5.3.2. Critical clock asynchronism

Each algorithms' sensitivity to clock asynchronism error is now computed. From the results presented above it is expected that MLR is more robust that PCA to surges in asynchronism error.

Simulations were run with the subsequent input variables, with results presented in Figure 23:

- Number of Clients: 100
- Number of Readings: 100, 300 and 500
- Number of runs per data point: 20
- Maximum absolute clock asynchronism: [0;120] seconds

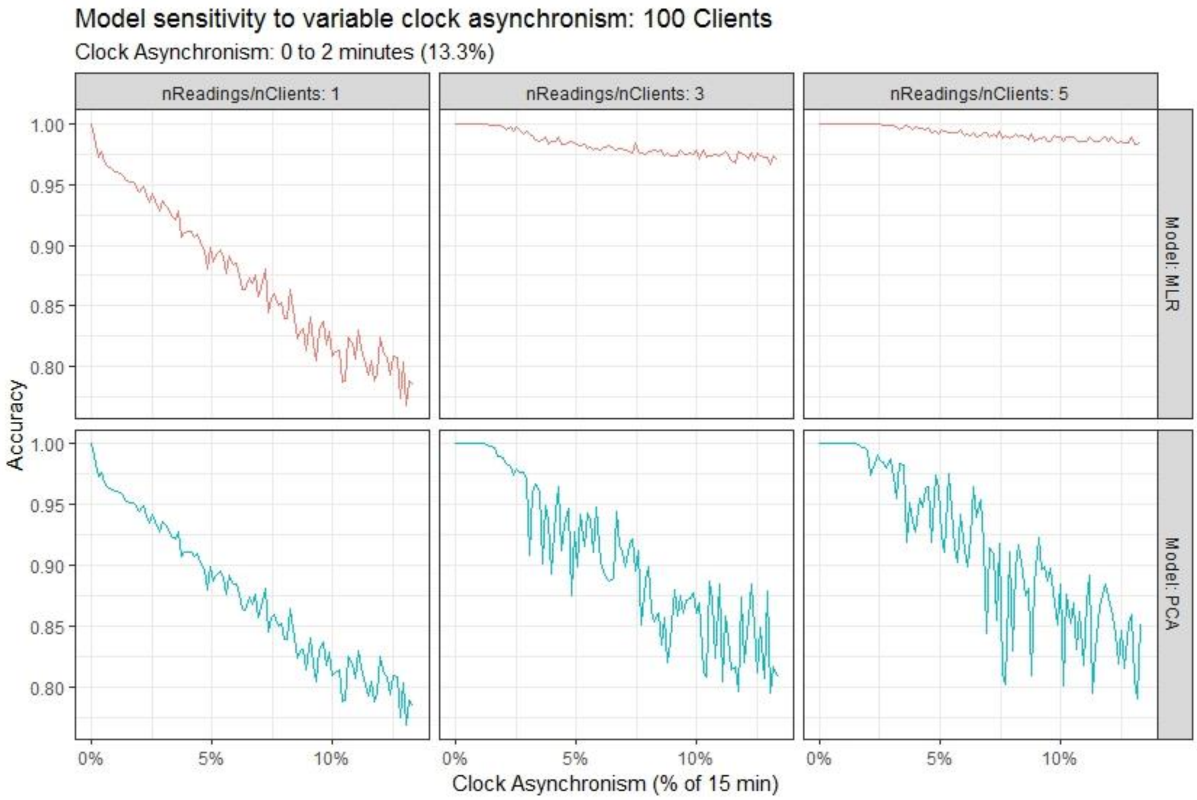

Figure 23 - Model sensitivity to clock asynchronism error

In fact, results are 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. Furthermore, bear in mind the results presented are the average of 20 runs and thus, if we plot a single run, results will even more intermittent for PCA.

Taking as an example the last data point in the above simulation, MLR achieves 98% accuracy for 500 readings, even for a maximum of 2 minutes of clock asynchronism, while PCA hovers around 85%.

Finally, results are offered for what will be considered as the critical error of maximum clock asynchronism of 1 minute. Input variables are as follows:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Maximum clock asynchronism error: 60 seconds

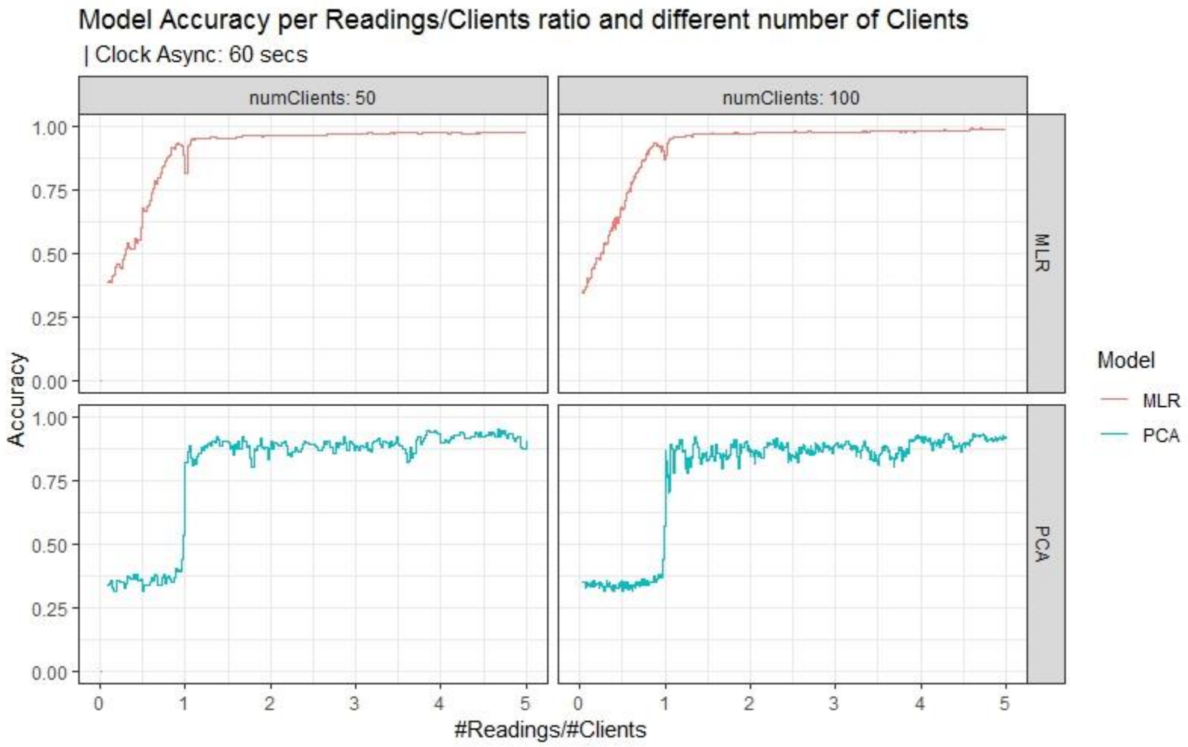

Figure 24 - Model accuracy with critical clock asynchronism error

Evidence demonstrates that MLR algorithm's accuracy improves with the number of given measurements, achieving approximately 100% accuracy for a ratio of 5 times the number of readings per number of clients. On the other hand, clock asynchronism error deteriorates PCA's performance, averaging only 90% accuracy.

5.4. Clock skew

5.4.1. Typical clock skew:

The second of clock errors is now presented. Considering clock skew errors, Figure 25 illustrates the results of applying a typical error as defined by the succeeding variables:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Clock skew error: 5%

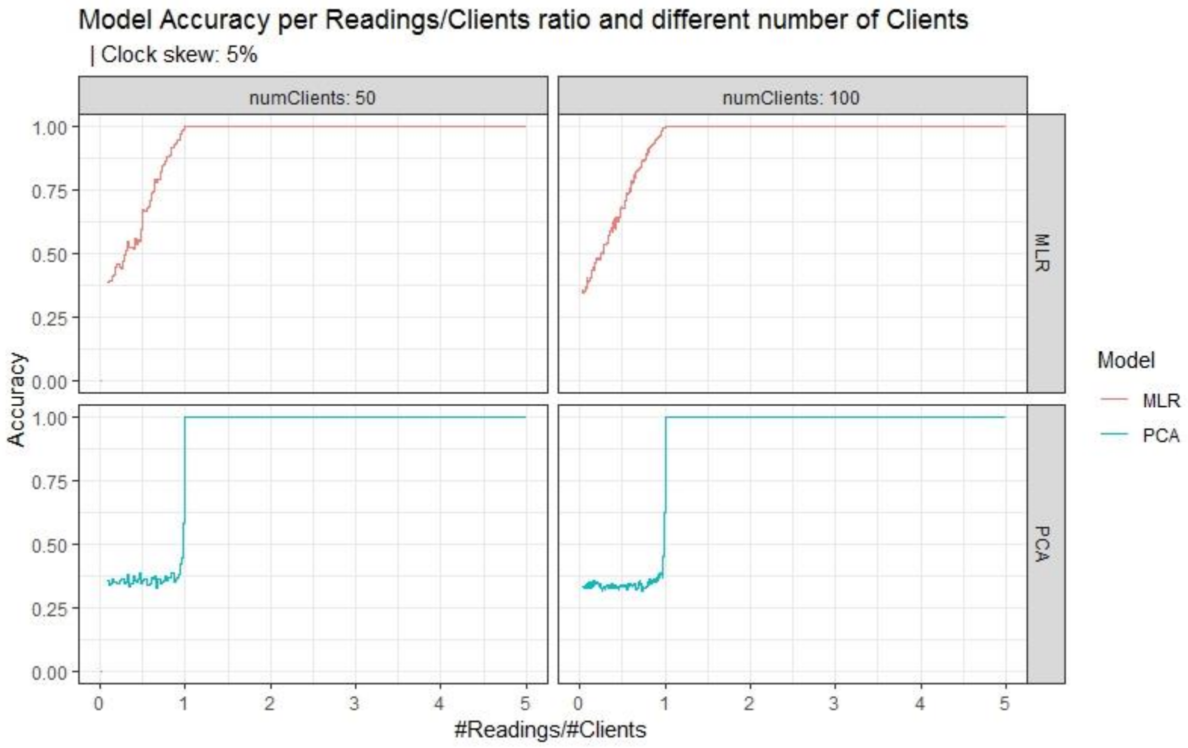

Figure 25 - Model accuracy with 5% clock skew error

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, at least when modelled as described in paragraph 4.3.3.

5.4.2. Critical clock skew

To confirm the previous proposition, each algorithms' sensitivity to increasing clock skew errors is now plotted, using the following inputs:

- Number of Clients: 100
- Number of Readings: 100, 300 and 500
- Number of runs per data point: 20
- Clock skew error: [0%;50%]

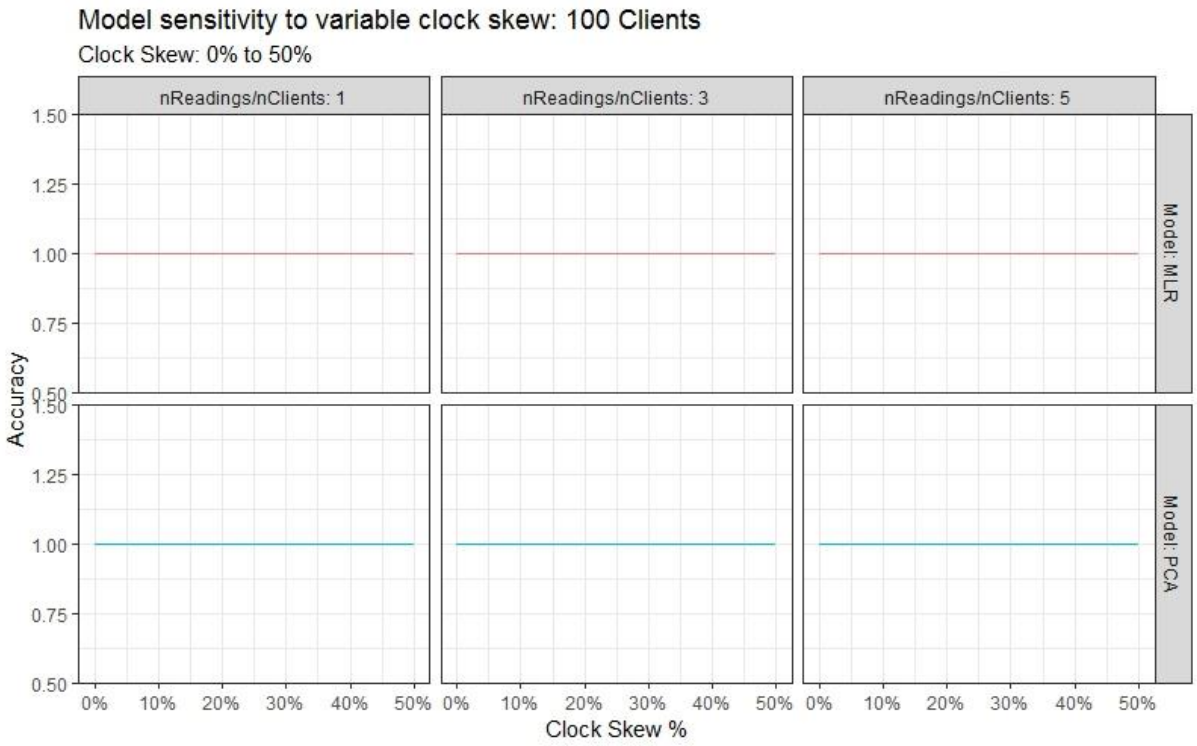

Figure 26 - Model sensitivity to clock skew

The results revealed in Figure 26 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. Consequently, there is no need to define a critical error for clock skew.

5.5. Copper losses

5.5.1. Typical 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 shown in Figure 27 for the given inputs:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Copper losses interval: [2%,10%]

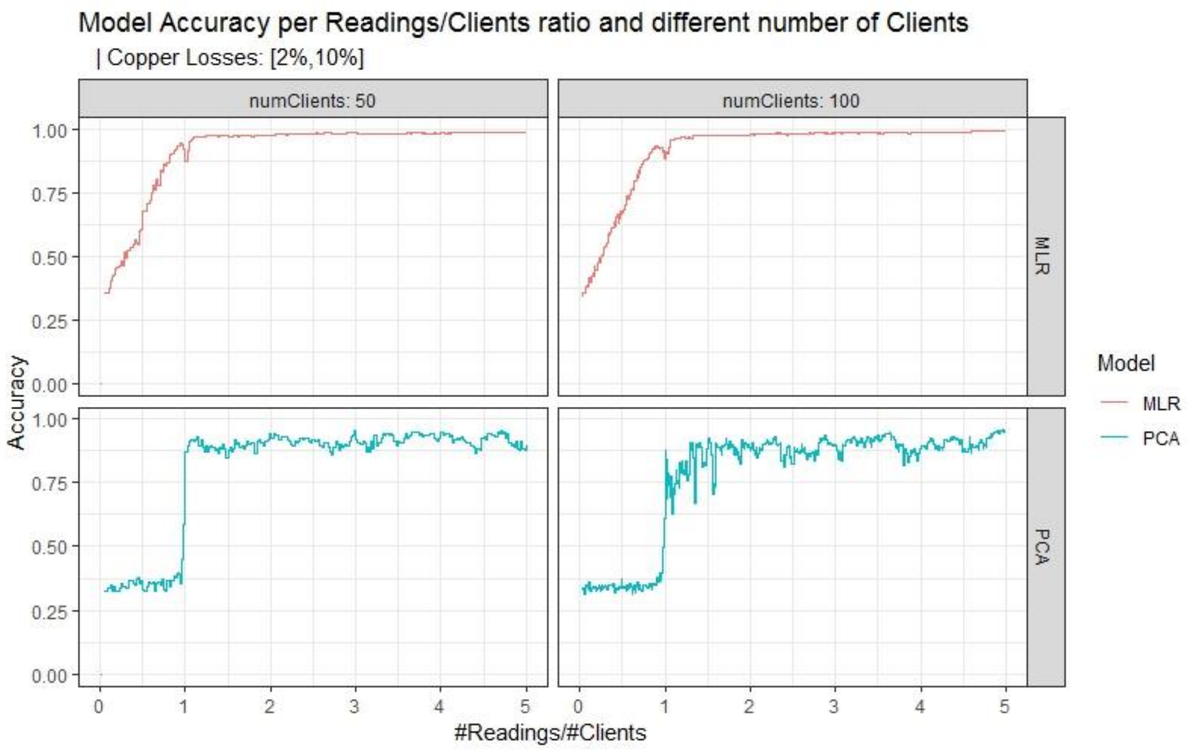

Figure 27 - Model accuracy with typical copper losses

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.

5.5.2. Critical copper losses

In order to determine a critical error interval for copper losses, Figure 28 plots each algorithms' sensitivity to an increase in both minimum and maximum copper losses with the additional input variables

- Number of Clients: 100
- Number of Readings: 100, 300 and 500
- Number of runs per data point: 20
- Minimum copper losses: [0%;10%]
- Maximum copper losses: [10%, 50%]

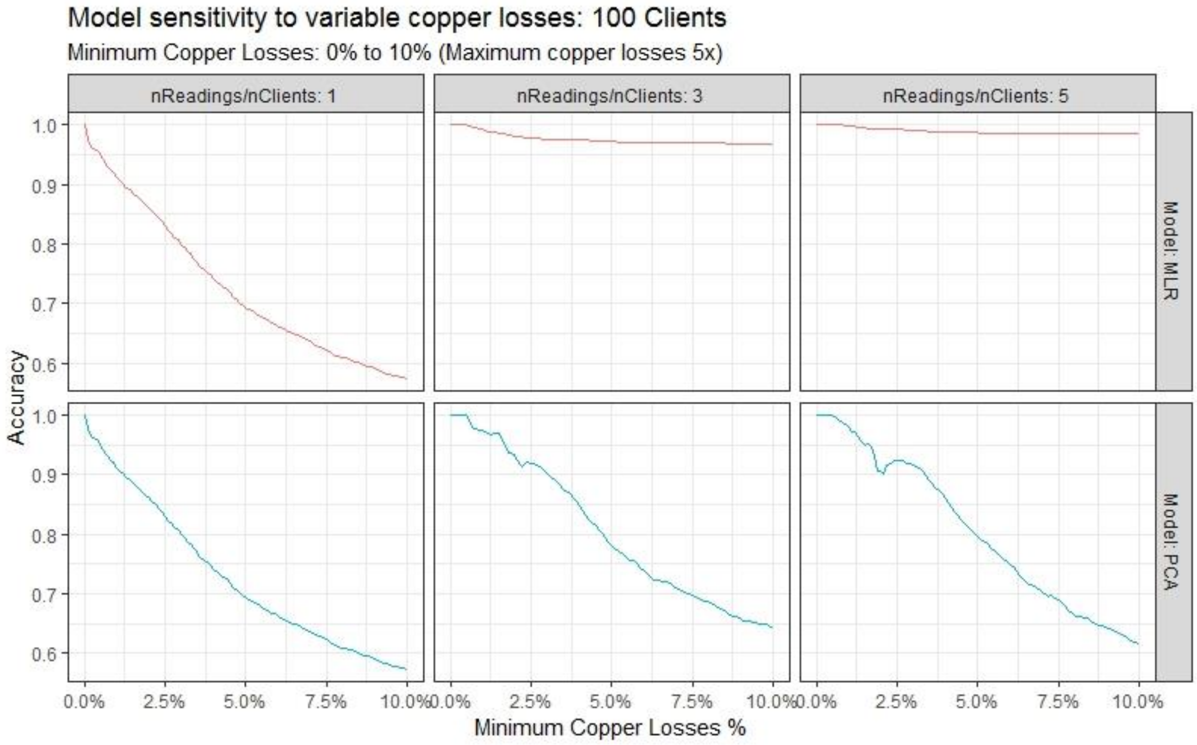

Figure 28 - Model sensitivity to copper losses

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 3.

In order to find the critical value for copper losses, let it be considered the minimum copper losses value that makes PCA achieve 90% accuracy when there are 3 times more readings than clients. The critical copper losses interval is thus from 3% to 15%.

Figure 29 below presents the results with the following input:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Copper losses interval: [3%,15%]

These results are consistent with the previous experiment with a typical error, showing a decrease in average accuracy for PCA algorithm while MLR's performance remains roughly the same.

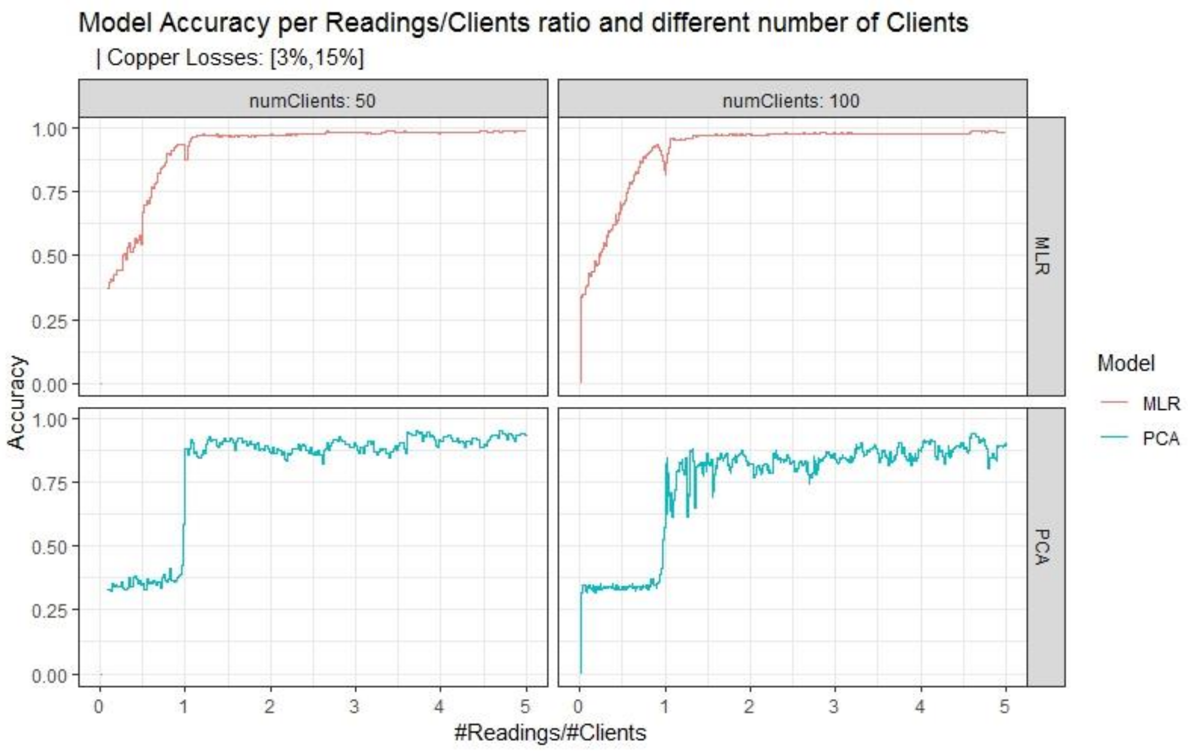

Figure 29 - Model accuracy with critical copper losses

5.6. Missing Clients

5.6.1. Typical missing clients

The final section on isolated errors presents the results for testing the data with missing clients. Algorithms' accuracy was computed with the following configuration:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Number of missing clients: 5

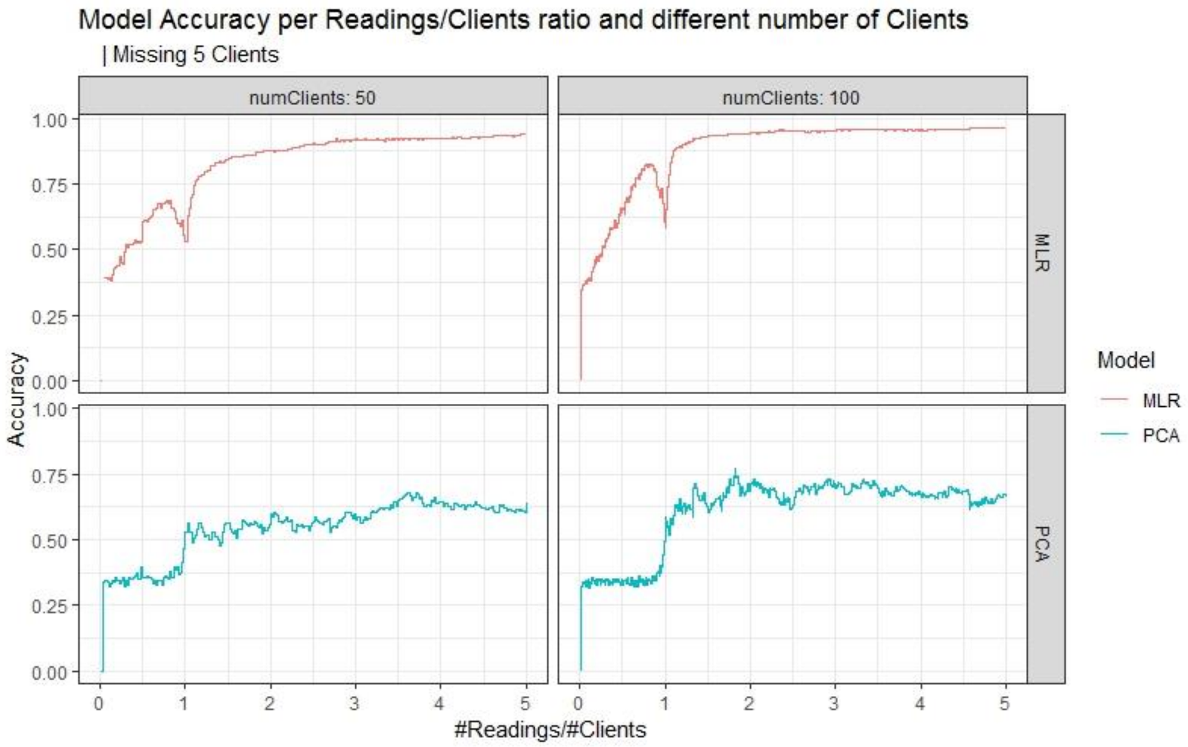

Figure 30 - Model accuracy with 5 missing clients

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 from nearly a destructive effect, hovering around 66% accuracy.

5.6.2. Critical missing clients

Figure 31 illustrates each model's sensitivity to an increasing number of missing clients, including the following inputs:

- Number of Clients: 100
- Number of Readings: 100, 300 and 500
- Number of runs per data point: 20
- Number of missing clients: [0, 20]

It is possible to observe that for each customer that has been 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.

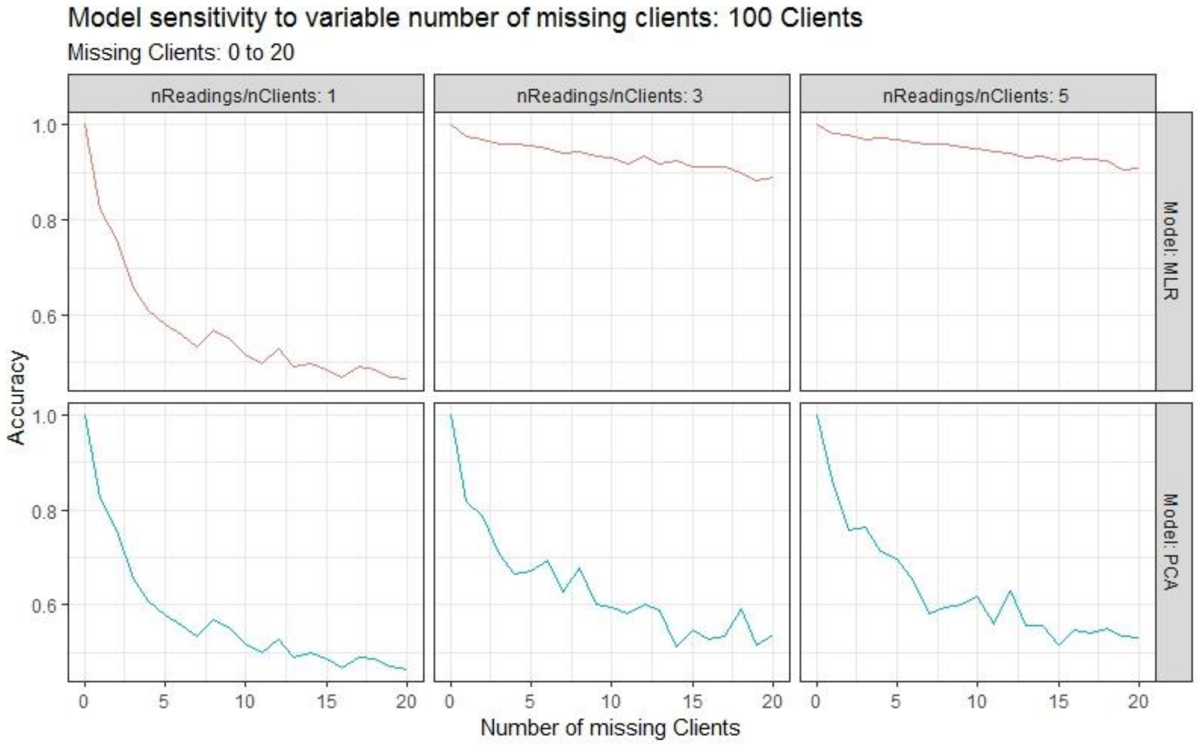

Figure 31 - Model sensitivity to variable number of missing clients

Let it be considered that 10 missing clients is the critical value for this type of error since it represents an inflection point in the curves plotted above, where the drop in model accuracy decelerates.

The following Figure 32 shows the performance of both algorithms accuracy when dealing with 10 missing clients' data. Simulations were run using the below-mentioned parameters:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Number of missing clients: 10

Once again, successive increases in missing data have a much harsher influence on PCA's performance when compared to MLR. In fact, given 50 clients, 10 missing clients corresponds to 20% of all clients' information being omitted and PCA's accuracy averages around 50%. Although 20% energy theft is exceedingly high, some 3rd word or developing countries are faced with a similar reality [37].

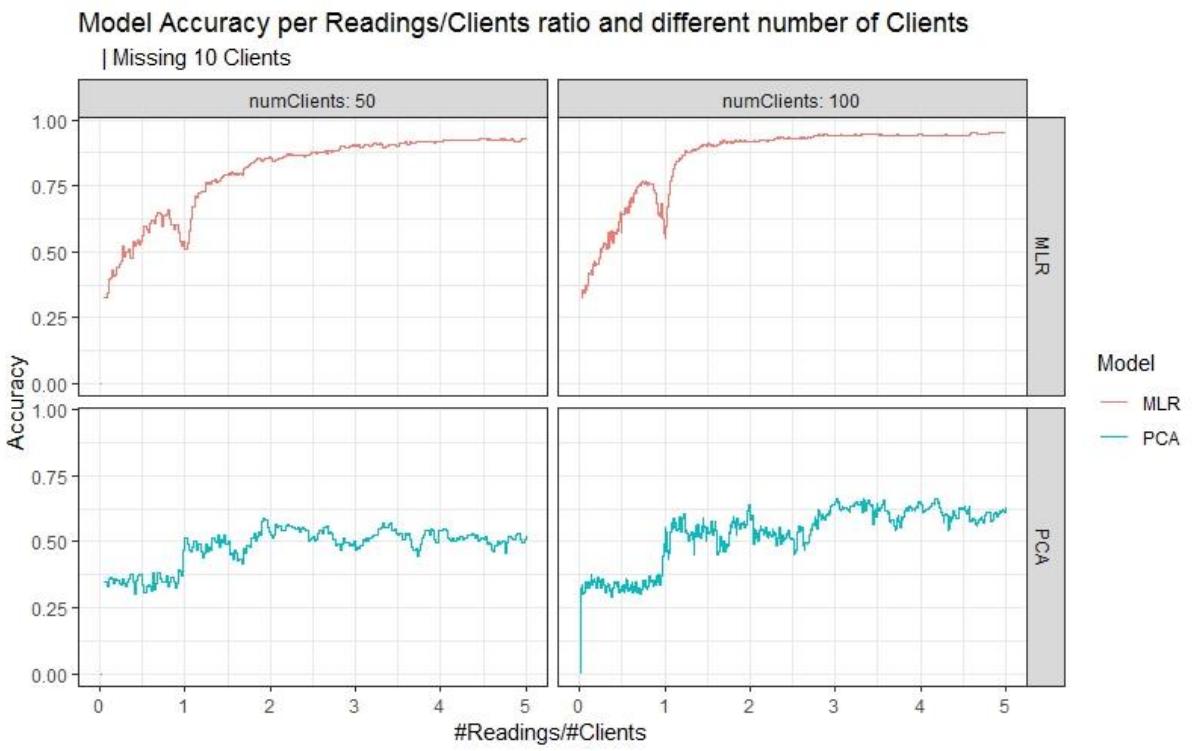

Figure 32 - Model accuracy for 10 missing clients

5.7. All errors simultaneously

5.7.1. All errors with typical values

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. It is possible to observe from Figure 33 that total errors are clearly visible even when only typical values, close to reality, are applied. The errors included in the simulation were the following:

- Number of clients: 100
- Meter class accuracy: 99.5%
- Clock asynchronism error: 45 seconds
- Clock skew error: 5%
- Copper losses: [2%, 10%]
- Missing clients: 5

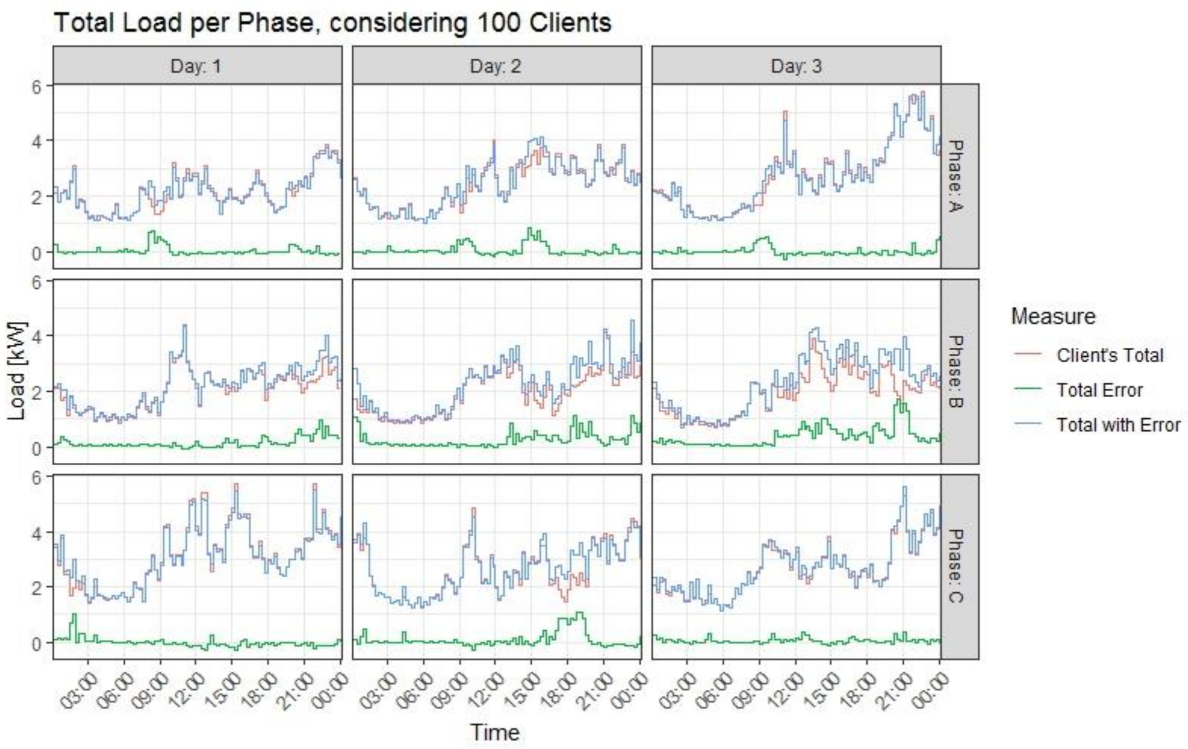

Figure 33 - Total load per phase with all typical errors

The ensuing Figure 34 plots each model's accuracy when all typical errors are included. Testing was done with the subsequent input variables:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- All typical errors included as per Figure 33

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 behaviour 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.

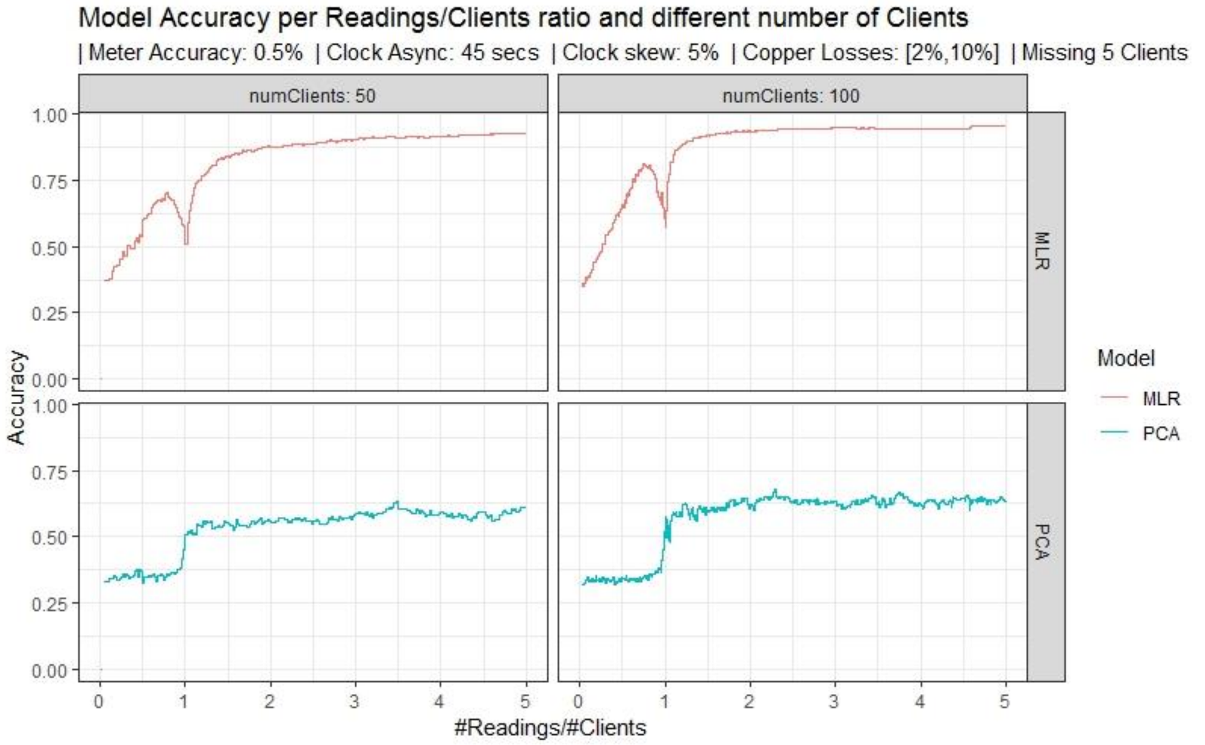

Figure 34 - Model accuracy with all typical errors

5.7.2. All errors with critical values

As a challenge to further assess the robustness of both algorithms, in particular to test the limits of MLR's performance, a simulation with all errors with critical error was performed.

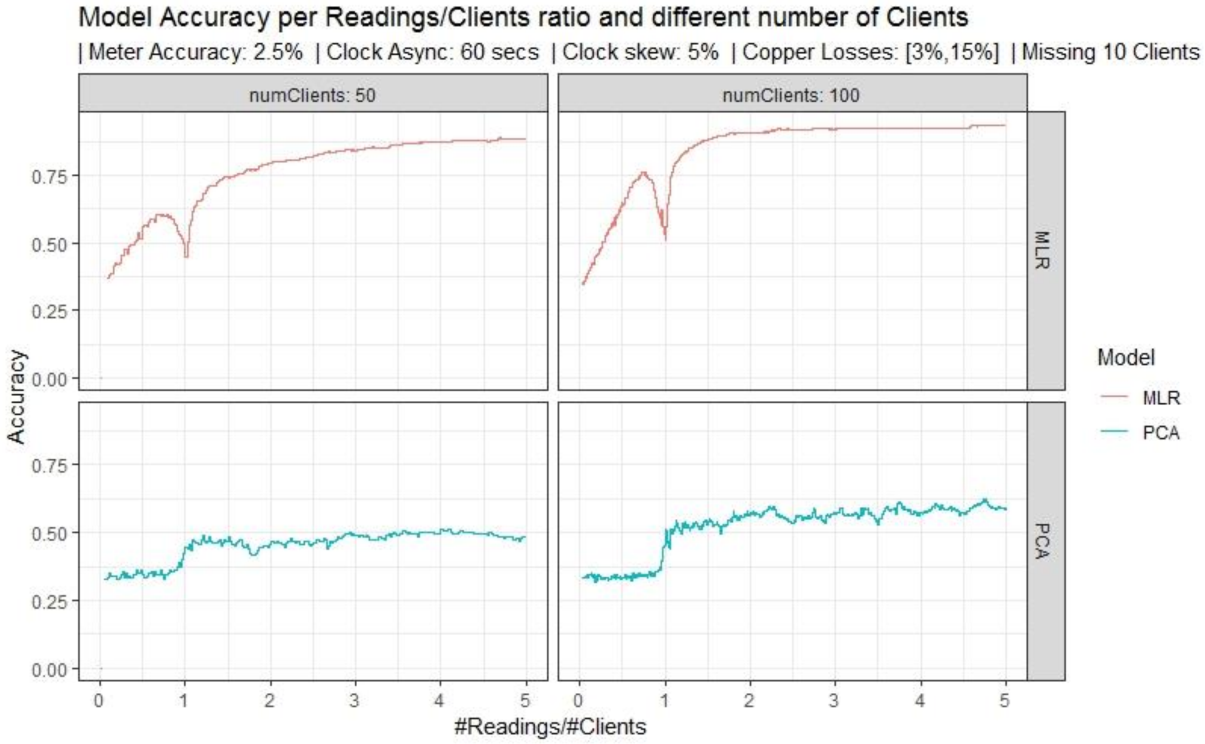

Figure 35 - Model accuracy with all critical errors

Figure 35 above has been computed with the following characterization:

- Number of Clients: 50 and 100
- Number of Readings: [0; 500]
- Number of runs per data point: 20
- Number of missing clients: 10
- Meter class accuracy: 97.5%
- Clock asynchronism error: 60 seconds
- Clock skew error: 5%
- Copper losses: [3%, 15%]
- Missing clients: 10

Extraordinarily, MLR still manages to output over 90% accuracy which further increases the confidence in this algorithm for inferring customer phase connectivity in the presence of multiple technical and commercial errors.

6. 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.

Bibliography

[1] J. P. Satya, N. Bhatt, R. Pasumarthy, and 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, and 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] V. Arya *et al.*, “Phase identification in smart grids,” *2011 IEEE Int. Conf. Smart Grid Commun.*, pp. 25–30, 2011.

[4] 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

[5] 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/

[6] Y. Liao, Y. Weng, M. Wu, and R. Rajagopal, “Distribution grid topology reconstruction: An information theoretic approach,” *2015 North Am. Power Symp. NAPS 2015*, 2015.

[7] W. Wang, N. Yu, B. Foggo, and J. Davis, “Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data,” in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 259–265.

[8] 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

[9] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, “Big data analytics in power distribution systems,” *2015 IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf.*, pp. 1–5, 2015.

[10] W. Wang, N. Yu, and Z. Lu, “Advanced Metering Infrastructure Data Driven Phase Identification in Smart Grid.” [Online]. Available: https://intra.ece.ucr.edu/~nyu/papers/2017-GREEN-PhaseID.

[11] D. K. Chembe, “Reduction of Power Losses Using Phase Load Balancing Method in Power Networks,” *Proc. World Congr. Eng. Comput. Sci.*, vol. I, 2009.

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

[13] S. J. Pappu, N. Bhatt, R. Pasumarthy, and 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.

[14] F. Olivier, D. Ernst, and F. Raphaël, “Automatic phase identification of smart meter measurement

data," *CIRED 2017 24th Int. Conf. Electr. Distrib.*, no. June, p. 4, 2017.

[15] Origo, "How to Choose a Phase Identification System," 2006. [Online]. Available: http://www.origocorp.com/Papers/PhaseID-How_to_Choose.pdf.

[16] C.-S. Chen, T.-T. Ku, and C.-H. Lin, "Design of phase identification system to support three-phase loading balance of distribution feeders," *IEEE Trans. Ind. Appl.*, no. 48(1), pp. 191–198, 2012.

[17] Z. S., M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, "Three-pase ac system impedance measurement unit (imu) using chirp signal injection," *Appl. Power Electron. Conf. Expo.*, 2013.

[18] S. Bandyopadhyay *et al.*, "Machine learning for inferring phase connectivity in distribution networks," *2015 IEEE Int. Conf. Smart Grid Commun. SmartGridComm 2015*, pp. 91–96, 2016.

[19] R. Mitra *et al.*, "Voltage Correlations in Smart Meter Data," *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15*, pp. 1999–2008, 2015.

[20] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," *52nd IEEE Conf. Decis. Control*, pp. 1659–1664, 2013.

[21] G. Cavraro, V. Kekatos, S. Member, and S. Veeramachaneni, "Transactions on Smart Grid Voltage Analytics for Power Distribution Network Topology Verification," *IEEE Trans. Smart Grid*, vol. 3053, no. c, pp. 1–10, 2017.

[22] J. D. Watson, J. Welch, and N. R. Watson, "Use of smart-meter data to determine distribution system topology," *J. Eng.*, pp. 1–8, 2016.

[23] B. Nicoletta, M. Davide, and M. Riccardo, "Topology Identification of Smart Microgrids." [Online]. Available: http://automatica.dei.unipd.it/tl_files/utenti/lucaschenato/Classes/PSC12_13/Projects/PSC13_relazione_TopologyID.pdf.

[24] T. A. Short, "Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 651–658, 2013.

[25] K. Soumalas, G. Messinis, and N. Hatziargyriou, "A data driven approach to distribution network topology identification," *2017 IEEE Manchester PowerTech*, pp. 1–6, 2017.

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

[27] S. Narasimhan and N. Bhatt, "Deconstructing principal component analysis using a data reconciliation perspective," *Comput. Chem. Eng.*, vol. 77, pp. 74–84, Jun. 2015.

[28] "RStudio Version 1.0.143." 2016.

[29] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodoligies," *Artif. Intell. Rev.*, vol. 22, no. 1969, pp. 85–126, 2004.

[30] S. (Australia) P. Ltd, "Electricity metering accuracy explained," 2014 [Online]. Available:

https://www.ecdonline.com.au/content/electrical-distribution/article/electricity-metering-accuracy-explained-372339275#axzz5YByMubtL

[31] Microsemi, “The New Role of Precise Timing in the Smart Grid The New Role of Precise Timing in the Smart Grid.” [Online] Available: https://www.microsemi.com/document-portal/doc_view/133267-the-new-role-of-precise-timing-in-the-smart-grid

[32] B. Ahuja, “Skew-free clock signal distribution network in a microprocessor of a computer,” in U.S. Patent No. 5,307,381 issued April 26, 1994.

[33] M. U. Hashmi and J. G. Priolkar, “Anti-theft energy metering for smart electrical distribution system,” *2015 Int. Conf. Ind. Instrum. Control. ICIC 2015*, no. July 2015, pp. 1424–1428, 2015.

[34] M. Clemence, R. Coccioni, and A. Glatigny, “How Utility Electrical Distribution Networks can Save Energy in the Smart Grid Era,” *Schneider Electr.*, 2013.

[35] S. Sahoo, D. Nikovski, T. Muso, and K. Tsuru, “Electricity theft detection using smart meter data,” *2015 IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf.*, no. June 2015, pp. 1–5, 2015.

[36] “Fighting Electricity Theft with Advanced Metering Infrastructure,” *ECI Telecom Ltd.*

[37] T. B. Smith, “Electricity theft: A comparative analysis,” *Energy Policy*, vol. 32, no. 18, pp. 2067–2076, 2004.