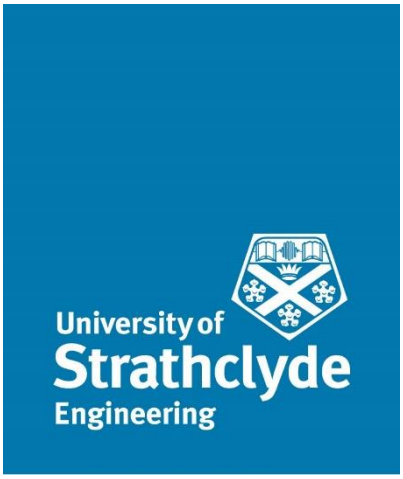

Department of Mechanical and Aerospace Engineering

Consumer Phase Identification for Different Scenarios of Low Voltage Networks with Constrained Smart Meter Data in the UK

Author: María Peñarrubia Pozo

Supervisor: Graeme Flett

A thesis submitted in partial fulfilment for the requirement of the degree

Master of Science

Sustainable Engineering: Renewable Energy Systems and the Environment

2019

Copyright Declaration

Signed: María Peñarrubia Pozo Date: 23/ 08/ 2019

Abstract

Phase imbalance is an increasingly important problem for low-voltage networks in the UK, mainly because they can lead to energy losses and lower hosting capacity of embedded technologies, such as PV-DG. This is of vital importance nowadays since PV-DG is precisely one of the main factors driving phase imbalance. In addition, considering that PV-DG is experiencing a growing penetration, if not tackled properly it might cause major issues in the operation and management of LV networks. For this reason, many methods have already been developed, which put forward phase identification approaches.

Phase identification methods allow identifying the phases to which customers (and/or feeders) are connected to. If network operators have this information, they could take actions to correct these imbalances, and avoid the unfavourable impacts of unbalanced networks.

The main drawback of the existing methods is that they are usually tested for scenarios where, for example, the SM uptake is very high, there is information at both the transformer and the consumer point, or there is special equipment used that directly allows identifying the phases of the SM. The real situation for network operators in the UK is currently very different from these scenarios, and it is not expected to change soon. Nowadays, the SM coverage is very low and even when SMs are installed in LV networks, the information sent to network operators is usually of low granularity and rather limited.

Therefore, the main objective of this thesis is to develop a profound analysis of the existing solutions, in order to test and determine if these methods could be adapted so that they could solve the phase identification problem in highly constrained scenarios.

Drawing upon the findings from the literature review, it was concluded that hierarchical clustering and k-means clustering techniques could be suitable for the study. These methods were firstly compared for different levels of SM coverage in order to select the most appropriate method.

It was found out that k-means clustering was more appropriate and following analyses in the thesis were focused on investigating how different variables affect their accuracy

when allocating phases. Some of the variables to be tested are the presence of PV on the rooftops, decreasing levels of SM penetration, and type of SM measurements.

In addition, in order to correct the main weakness of k-means clustering, an improved algorithm was presented. This weakness is the need to define the number of clusters to be formed, or in this case, the number of different phases to be identified in a SMs data set.

With this improved k-means approach several scenarios with moderate PV-DG levels and SM penetration were tested, resulting in very promising results. Conclusions on the best type of voltage measurements were also derived from these analyses.

Acknowledgements

In the first place, I would like to thank my supervisor Graeme Flett, who was very encouraging and helpful thought the whole project, even though I decided to take a different direction from the original one. Also, I would like to thank all the lecturers, and especially, Paul Tuohy who has been an excellent master's coordinator

I would also like to thank Scottish Power for providing me with data and help to develop this project. In particular, thank you to the people from the Energy Networks department, Fiona Fulton, John Gray, and special thanks to Maizura Mokhtar who was always willing to meet and solve my questions. I would also like to mention that I am genuinely grateful to Iberdrola for grating me the 'Scholarship for master's studies' which enabled me to study this master.

Finally, I would like to thank my family, special thanks to my brother Alberto, who helped me a lot to get through the last months of the thesis. Also, thank you to my friends and my boyfriend, Daniel, who despite the distance were an unlimited resource of support.

Thank you so much to all the lovely people I met in Glasgow, who are too many to mention all. However, I would like to give special thanks to Evie, Olatz, Naia and Fernando for all the hours spent together and putting up with me even in the most challenging times.

Abbreviations

DNO	Distribution Network Operator
EV	Electric Vehicle
GB	Great Britain
GW	Gigawatts
LCT	Low Carbon Technology
LTC	Load Tap Changer
LV	Low Voltage
PLC	Power Line Carrier
PV	Photovoltaic panel
PV-DG	Solar photovoltaic distributed generation
SM	Smart Meter
THD	Total Harmonic Distortion
TSO	Transmission System Operator
UK	United Kingdom

List of Figures

Table of Contents

1 Overview

1.1 Introduction and Motivation

UK government has set out a number of incentives and grants to boost the deployment of distributed generation (DG) and electrification of transport, with the goal of achieving the targets on carbon emissions established for 2050 (Department of Energy & Climate Change, 2011). While this is very beneficial for climate change and carbon emission reductions, it will also increase the demand for electricity, and the number of devices connected to the LV network.

It therefore becomes apparent that electrification of transport, embedding DG in the LV, increasing participation of users, and increase of monitoring and active management of LV networks, will pose fundamental challenges in the years to come, as well as, fantastic opportunities to make a difference and evolve in the way we consume and produce energy. We are therefore moving towards different energy landscapes, and it looks like an essential part of the change is going to start from the LV networks.

In the past, the LV network, which is going to be the main focus of this thesis, did not require to be monitored in detail, and the focus was set on the transmission networks instead. However, due to rollout of these new agents in the LV network and the widespread of smart grid and smart meters, this ‘fit-and-forget’ approach is no longer suitable for distribution networks (Mokhtar, et al., 2019).

One of the main issues of these LV networks is the phase identification problem. As already mentioned, there is a significant amount of existing methods in literature. Nonetheless, it should be noted that the suitability of these methods is dependent on the data and equipment available for the networks under study. Many of these methods requite full observability of the network, with high levels of penetration of SMs, and sometimes even the installation of additional equipment (such as microsynphasors, or more advanced SM models). Nevertheless, assuming that this level of data detail is going to be always available in real life is not realistic. What is more, SM roll-out can sometimes be troublesome, and high levels of penetration in LV networks are not likely at this moment. In fact, the current level of penetration of SMs is considerably low, approximately a 25% per cent in GB by the end of June 2018 (Department for Business, Energy & Industrial Strategy, 2019).

In light of the above, it is important to investigate the feasibility of other methods in tackling the phase identification problem, so that they can be accurate even in scenarios where there is missing information due to the low coverage (or poor operability) of SMs in LV networks.

Several real LV networks will be analysed in this thesis, which present a limited number of SMs, as well, as constrained SM measurements. The main reason for using these highly constrained networks is that these networks are intended to represent the current situation of UK distribution networks. Nowadays, the level of SM uptake in the UK is rather low, and therefore, network operators have limited data on LV networks, which lead to unbalanced systems, and thus, to an inefficient operation and energy losses. Not that most distribution networks are unbalanced, but the current little monitoring does not allow to take actions to tackle the situation. Moreover, this situation will be compounded by the increasing adoption of DG in the LV networks. Unbalanced networks, in turn, reduce the hosting capacity of DG, therefore it is of great interest to have balanced LV networks, if striving to achieve scenarios where there is a more significant presence of these LCT.

For the above reasons, in this project, firstly a thorough review of existing methods was carried out in order to fully understand how the phase identification problem could be solved, given the constraints and characteristic of the LV networks under study. The main finding from research led to choose two preliminary machine learning techniques, which were applied to the test networks, in order to determine if there is any method that can effectively solve the phase identification problem for the given case study.

After testing the two preliminary methods, only one of them demonstrated to yield accurate results for the phase matching, and it was further tested to determine how different factors, such as lower levels of SM penetration, or the moderate PV penetration scenarios affected the accuracy of its phase allocation.

Automatic phase identification methods, such as, the proposed approaches in this thesis, are the precursor to intelligent automated systems that will enable to monitor the network with a high level of granularity and detect and rectify unbalances for optimal performance. In addition to enable higher levels of PV panels installed on the rooftops.

1.2 Aims and Objectives:

The main objective of this thesis is to investigate if phase identification methods can be successfully applied for highly constrained LV networks, such as the networks under study. The characteristics of this networks and its limitation are further detailed in section 5, but it should be noted that the type and size of data set are the main factors that determine the suitability of phase identification methods. Consequently, such highly constrained scenarios pose an important challenge, since none of the methods found in literature were tested for similar limited data sets.

Consequently, the following are the main goals of the project:

- To carry out an extensive review of already existing phase identification approaches, in order to gain a deep understanding of the potential solutions for the test networks. From this research process, it is, therefore, intended to propose which methods are theoretically suitable for the case studies.

- These proposed methods will be tested for the test networks, in order to check the can also be successfully applied to the case studies of this thesis. They will also be compared, with the objective of selecting the most effective method, and later perform a series of analyses.

- The definitive methods will be tested for different scenarios, that is, it will be checked how the accuracy of the phase matching is affected by different levels of SM penetration, and for scenarios both without PV-DG and with a moderate PV uptake. This will be done, in order to be able to conclude under what conditions high levels of accuracy can be achieved for the final phase identification method.

1.3 Project Outline

The first section provides an overview of the current situation of DG and SMs in the UK. It opens by outlining the main principles of LV network in the UK, as well as, its current state. Next, it is addressed the increasing levels of penetration of DG in distributed networks. It is important to understand how the DG is evolving and moving forward to an increase in its presence in LV networks. Lastly, the main facts about how SMs are being deployed in the UK are given, in order to understand, the current level of monitoring of LV networks.

Chapter 3 addresses what the phase identification problem consists of. It also details what phase imbalances are, which are the main factors that can cause such imbalances, as well as, the impacts of unbalanced network, and why it is so important to avoid phase imbalances by implementing phase identification methods.

The following section outlines relevant knowledge on phase identification methods that have already been proven to succeed in literature. This revision allows to assess and pre-select the most suitable methods for the case studies analysed in this thesis.

The next chapter provides detailed information on the proposed methods that will be applied to the test data sets, to determine their suitability and effectiveness. This chapter enables to understand how these methods rational work, in addition to information about how the Machine Learning Toolbox of Matlab was used to implement these methods.

Section 6 provides detailed information about the main characteristic and topology of the six LV networks to be studied. This information is important in order to understand the constraints and limitations of the available data. Existing methods for solving the phase identification problem are data-sensitive, therefore, this limitation will determine the suitability of the proposed methods.

Chapter 7 explains how the methodology process was developed in order to implement the proposed methods presented in section 5, and the different analyses that were performed to test how different scenarios and factors affect the accuracy of the phase allocation of these methods.

Section 8, 9 and 10 present the results obtained for these analyses, as well as, the discussion on the results obtained in each analysis.

Finally, the main finding and limitations of the whole project are presented in the conclusions section. In addition, in chapter 12, future lines of work, explore potential improvements and expansion of the studies carried out in this thesis.

2 Distribution Systems and SMs in the UK

This section provides an overview of the current state of distribution systems and smart meters in the UK. It gives a better understanding of the significance of this thesis for the future development of new energy landscapes.

2.1 Overview

Traditionally, the transmission (high voltage) system has always been analysed in more detail than distribution systems. In fact, it is common practice to monitor transmission lines every second and sometimes using complex algorithms. This is due to the fact that in the past the whole energy production tended to be centralised in big power plants, and this energy was distributed to substations near the consumption points through high voltage lines. Therefore, the main focus was placed on avoiding losses in the transmission lines, and it is the mission of Transmission system operators (TSOs) to take charge of monitoring and balance power flows to avoid them. (Dang-Ha, et al., 2015)

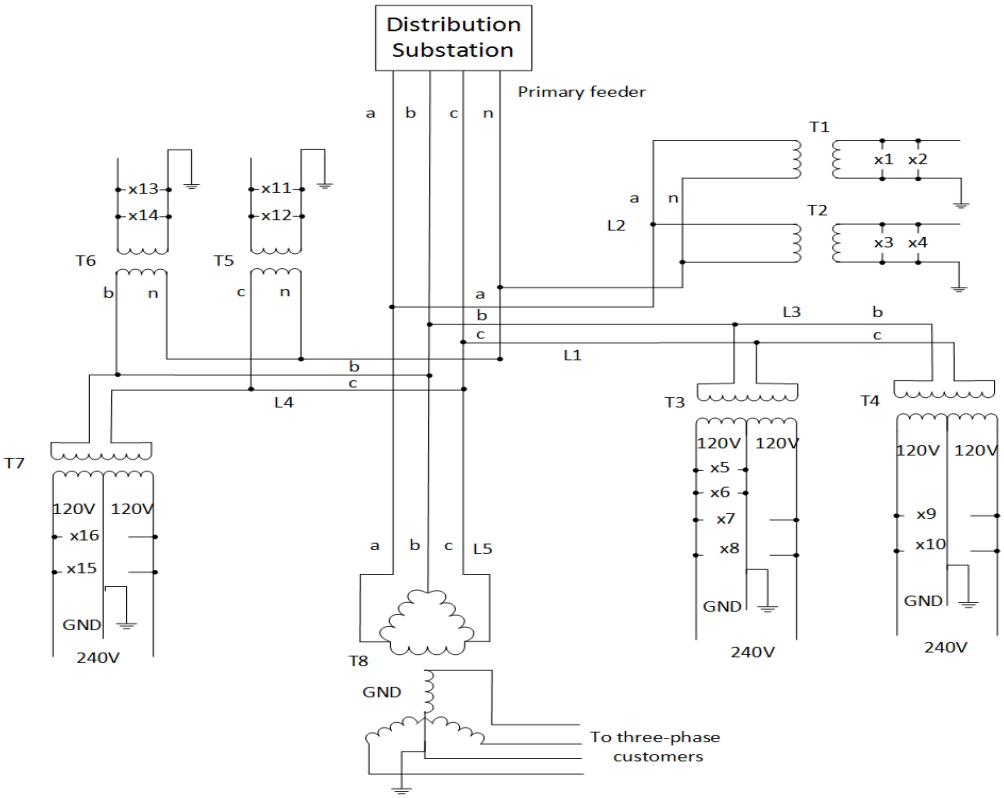

Figure 2.1 - Example of a distribution architecture in the UK (Wang, et al., 2016)

On the other hand, distribution systems structure is usually tree-like, also named 'distribution feeder'. As shown in Figure 2.1, a typical distribution system[1] the energy is transmitted from the transformer to the all the customers through the branches. (Morgan, et al., 2019). Unlike transmission lines, LV networks usually rely upon little control or monitoring. Consequently, this may pose a problem in future energy scenarios where electrification of heat and embedded energy resources (also known as Low Carbon Technologies, LCT) have a greater presence in distribution systems.

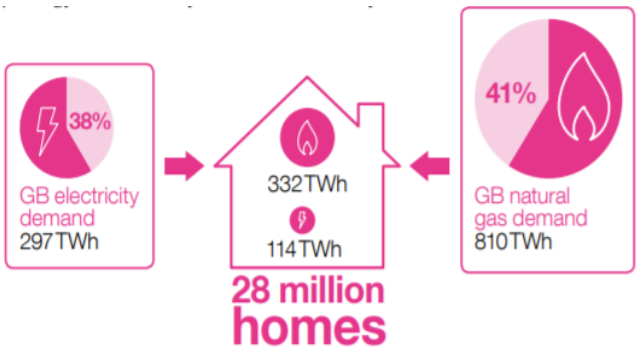

Figure 2.2 - Share of energy consumed by households in Great Britain (National Grid, 2018)

In addition to this, as shown in Figure 2.2, domestic energy consumption in Great Britain accounts for 38% of the total, and therefore, any problems caused in domestic LV networks may trigger significant losses for the whole energy network.

2.2 Distributed Energy Production in LV networks

In the UK, almost every domestic consumer is connected to the LV network (that is, the distribution network) and usually to single-phase cables (Mokhtar, et al., 2019). Therefore, for the deployment of residential LCT, they will mostly be connected as well to the single-phase LV supply trough 'line-to-neutral' cables. This will result in an increase of the resources (or loads) present in the network, and it is very likely that it would lead to imbalances, among other problems, and will certainly affect the operation of the LV networks. (Li & Crossley, 2014).

[1] In Figure 2.1, letters a, b, and c stand for the three phases. L represents a lateral, and T, the transformers, and x denotes consumers (Wang, et al., 2016)

Moreover, Distribution Network Operators (DNOs) have traditionally administrated and operated distribution networks passively, with unidirectional power flows. Since, for example, in domestic LV networks, customers appliances were the only loads in the system. However, with the increasing roll-out of LCT, this will no longer be possible because reverse power flow issues will be more and more frequent (Cipcigan & Taylor, 2007).

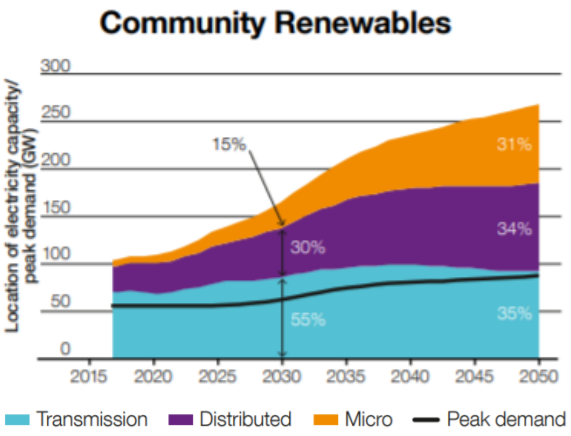

Figure 2.3 - Community Renewables scenario for the increase of the electricity peak demand in Great Britain (National Grid, 2018)

Moreover, Figure 2.3 - Community Renewables scenario for the increase of the electricity peak demand in Great Britain depicts the expectations regarding the increase of electricity peak demand estimated by (National Grid, 2018) in their 'Community Renewable Scenario', which considers a more decentralised energy landscape that could allow the achievement of 2050 targets.

For all the above reasons, higher levels of observability and active management of LV networks is becoming increasingly important; and this, in turn, is resulting in more emphasis on the role of DNOs (Hollingworth, et al., 2012).

It is also useful to highlight that these stated consequences of DG are caused by all types of embedded energy technologies, and not necessarily LCT (e.g. reciprocating engines). Nonetheless, renewable technologies, such as PVs or wind energy, due to their intermittent nature, constitute a particularly significant challenge for the optimal operation and management of distribution networks (Begovic, et al., 2012).

In this project, however, as already mentioned, the focus is going to be set on the impact of PV-DG in LV networks, and its relation to the phase identification problem. This is why the residential PV roll-out and its impacts are discussed in more detail below.

2.3 Domestic PV-DG and its Impacts on the LV Network

2.3.1 Domestic PV roll-out

It noteworthy how energy generation is being increasingly generated in a wide variety of decentralised ways, but nowadays, PV-DG is the most prominent (Hoornaert, et al., 2016).

Since 2010, as it can be observed in Figure 2, the PV installed capacity has increased up to exceeding the total capacity of 13 GW. While in Figure 2.4, it can be observed the UK government incentives that have fostered this rapid growth in recent years.

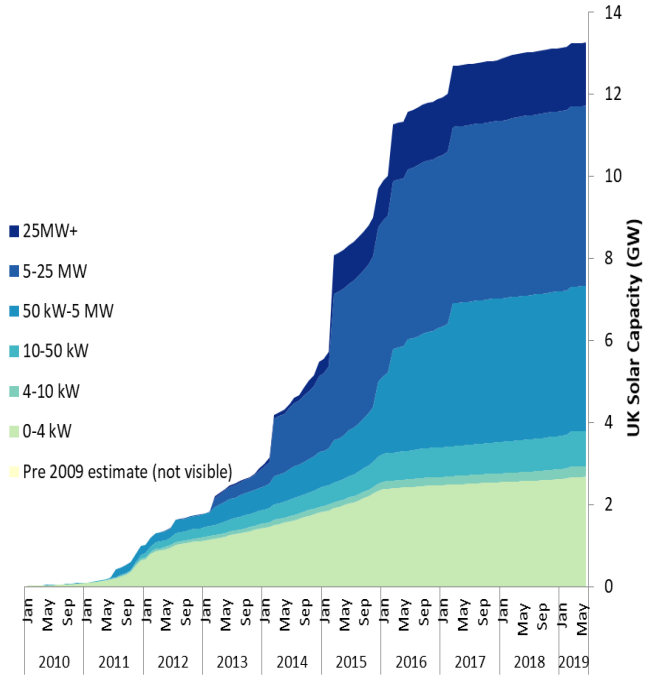

Figure 2.4 - UK solar deployment by capacity updated monthly (Department for Business, Energy & Industrial Strategy, 2019)

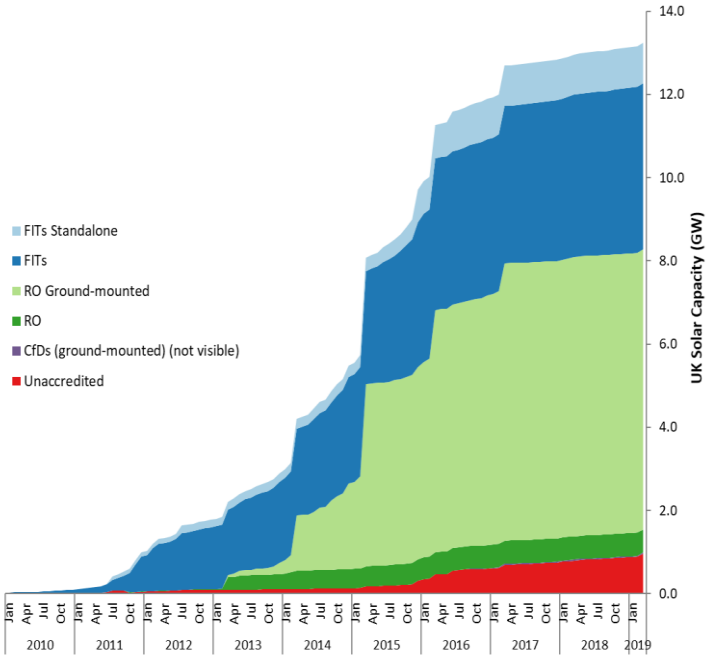

Figure 2.4 - UK solar deployment by accreditation updated quarterly (Department for Business, Energy & Industrial Strategy, 2019)

It is worth noting that around 20% of the total installed capacity comes from domestic PV-DG connected to the LV network (Department for Business, Energy & Industrial Strategy, 2019). This only reaffirms the importance of maximising the operation and management of PV-DG in the UK.

2.3.2 Impact of PV- DG on LV Networks

Even nowadays, PV-DG deployed in the LV networks has caused problems to the DNOs in the UK. For this reason, it is important to further investigate their effects and how to tackle this issue, so that future scenarios with higher degrees of domestic PV roll-out will not pose a problem for the network control (Navarro, et al., 2013).

Next are detailed some of the main impacts caused by PV-DG in LV networks (Begovic, et al., 2012) (Romero Agüero & Steffel, 2011):

1. Reverse power flow:

High degrees of PV-DG penetration can produce reverse active power flows at feeder and substation transformer level. This can be troublesome for overcurrent protection coordination, in addition to the operation of line voltage regulators.

2. Voltage Rise:

The integration of domestic PV-DG leads to changes in the feeder voltage profiles, as well as, voltage rises in the proximities of the PV site. This becomes especially evident when large PV-DG are deployed in rural areas since they tend to have long, lightly loaded feeders. Voltage rises can cause voltage violations on utility planning and industry standards, which in turn, may lead to customer dissatisfaction, and the need for using overvoltage protection systems of the PV-DG systems.

3. Voltage fluctuations:

Due to the intermittency of the PV generation, and since weather variability affects the efficiency, the output generation of PVs can be affected. Also, fluctuations in the feeder's voltage may happen at the distribution level. This phenomenon is more common on 'weak' feeders, that is, those that have short-circuit power levels, as well as, on sites where the PVs are located far from the transformer.

4. Interaction with voltage-controlled capacitor banks, an overvoltage protection operation of Load Tap Changers (LTCs), and line voltage regulators:

The two previous impacts, voltage rises and fluctuations, might, in turn, require the use of equipment and/or techniques to control them. The most commonly used ones are voltage-controlled capacitor banks, LTCs, and line voltage regulators. The overuse of these measures, however, can lead to increased maintenance costs, and eventually, cause augmented voltage fluctuations. Besides, repeated utilisation of capacitor banks can also lead to reverse power fluctuations.

5. Feeder loading and power decrease:

Feeder loading may be caused by low or intermediate levels of PV-DG penetration, and it is especially significant when the peak demand occurs during the daylight hours. While power decrease is usually a consequence of the reverse flow caused by larger amounts of PV-DG. Reverse power flows can lead to an increase in the distribution line and equipment loading, resulting in higher distribution line losses, and hence increase the feeder power losses. It should be highlighted that these losses can also be intensified if there are voltage rises in the network, due to the fact that distribution transformer core losses are proportional to nodal voltage.

6. Total Harmonic Distortion (THD) increase:

High penetration levels may trigger an increase of THD, which is gaining importance owing to the advent of electronic interfaced equipment, such as electric vehicle charging points or distributed energy storage. THD increase means that the presence and interplay among these devices connected to the network plus capacitor banks action can produce harmonic related problems like resonance.

7. Voltage and current unbalance:

The increasing presence of PV-DG connected to single-phase cables in the LV network might offset feeder currents, resulting in voltage and current unbalance.

It is worth noting that the phase identification problem in LV, which is the focus of the project, is intended to avoid unbalances and is further discussed in section 3.

2.4 SM Rollout in the UK

Smart meters are devices that allow to obtain advanced information about the network and enable two-way communication. In the UK, SM settings are programmed so that each half-and hour communicate their readings with energy suppliers and network companies (National Audit Office (NAO), 2018).

It is also noteworthy the governmental rollout programmes and statistics refer to both electric and gas meters, although in this thesis the spotlight is placed on electric SMs. Also note that, the following figures shown account for the whole Great Britain (GB), and not only the UK.

SMs are expected to play a crucial role in future energy scenarios, since accurate and complete information will be required for high levels of DG rollouts and the proliferation

of EVs. Accordingly, the UK Government presented ambitious plans to boost SM deployment, whose progress is discussed in the following subsections:

Rollout Plans Made by the Government:

- Back in 2008, the government made public its intention to promote the installation of SMs. It was also decided that the government will intervene to ensure the fulfilment of standards, and the operability of SMs (National Audit Office (NAO), 2018).

- In March 2011, a strategy for high-level implementation of SMs was put forward, which entailed two stages: Firstly, smart meters of first-generation (SMETS1) would be installed. Then, the second stage was supposed to take place in June 2014, after the DCC went live, and SMETS1 would be replaced with second generations SMs (SMETS2)[2]. Also, customers that still had conventional meter would replace them with SMETS2 (Department for Business, Energy & Industrial Strategy, 2019).

- In 2012, to make sure that if customers decided to switch supplier, the SMs would continue working, the government decided to establish new standards. These standards stated how SMs were supposed to be operated and connected to the central data and communications infrastructure (Data and Communications Company, DCC). Besides, customers were encouraged to start installing SMs even before the infrastructure was up and running.

- Also, in November 2012, the Department of Energy & Climate Change (hereafter referred to as 'the Department') legally obliged energy suppliers to have taken all the reasonable steps so that there would be full coverage of SMs by 2019. However, in 2013, this target was postponed until 2020 (National Audit Office (NAO), 2018).

[2] Note that there is no main difference in functionality between SMETS1 and SMETS2. But their differentiation lies in the way they communicate with the DCC, which is the main cause of problems in the installation of SM, and the reason why SMETS1 operation is being somewhat troublesome (National Audit Office (NAO), 2018).

2.4.1 Actual progress in the SM rollout:

- SMETS2s were supposed to start to be installed in June 2014. However, it was not until July 2017 that they were first introduced, and their installation process happened to be slower than expected by the Department. Also, it is unclear the number of years that it would take until the complete SMETS2 system works.

- Besides, due to the delay of SMETS2 installation, over 7 million more SMETS1 were installed than the originally planned figure. And although SMETS1 enabled customers to enjoy the benefits of smart metering, this considerable figure poses an important challenge for the correct development of the programme.

- The department has not yet tried to enrol SMETS1 into the DCC infrastructure, and it is not clear how feasible this task would be, resulting in the postponing of the deadlines for doing this.

- Another problem related to the operation of SMETS1 is that around 70% of them 'go dumb' when switching suppliers, meaning they lose their smart functionality (National Audit Office (NAO), 2018).

- To date, approximately 15.97 million domestic (both domestic and electric) SMs have been installed by large and small suppliers, which means a 6.9% of increase respect to previous quarter cumulative total figure (Department for Business, Energy & Industrial Strategy, 2019).

- Figure 2.5 shows the quarterly progress in the installation of domestic SMs by large suppliers. Even though it can be observed that in the last quarter of 2019 the number has decreased, the level of installations per quarter continues to be higher than 1 million for the ninth consecutive quarter (Department for Business, Energy & Industrial Strategy, 2019)

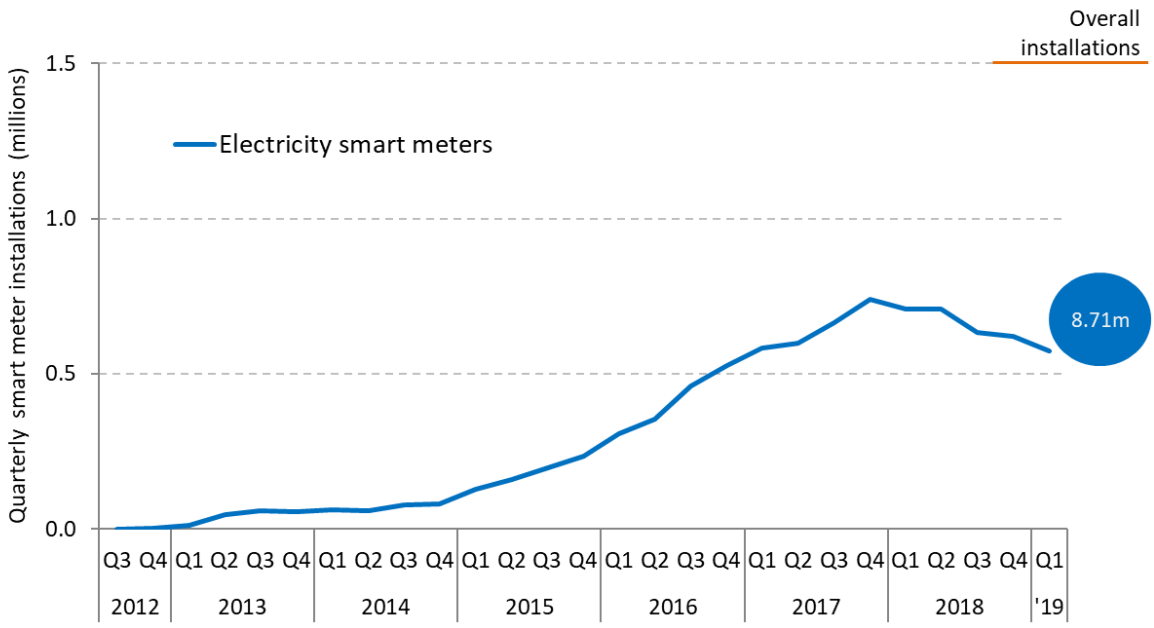

Figure 2.5 - Domestic SM quarterly installation for large energy suppliers (Department for Business, Energy & Industrial Strategy, 2019)

- Also, in Figure 2.6, it is portrayed the figures for the already installed SMs (by large suppliers) and the high number of traditional meters that have not yet been replaced. Note that only SMs (and not smart-type meters) are accounted as SMs in this case. In fact, by the end of June 2018, the penetration of SM was 25 per cent.

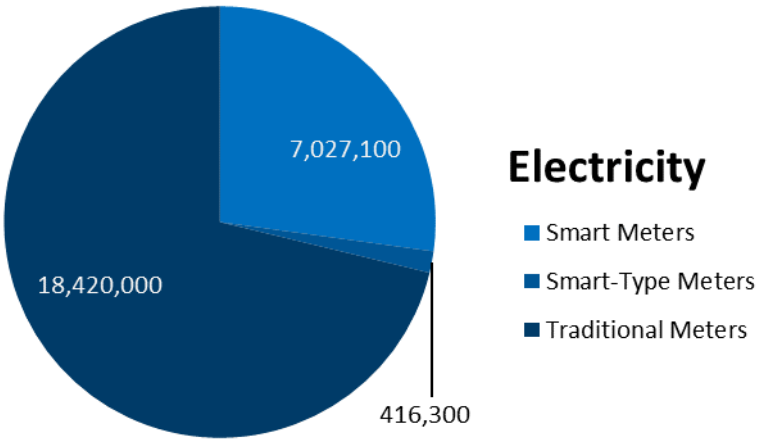

Figure 2.6 - Number of SM operated by large energy suppliers (Department for Business, Energy & Industrial Strategy, 2019)

- The NAO report concludes that the initial target of having an 80 per cent penetration by 2020 will very likely fall short, and the penetration is now expected to be around 70-75%.

From this subsection can be therefore concluded that a scenario where there is a 100% of SM in LV networks is not very likely to happen in the short term. This is precisely one of the primary motivations for the study developed in this project since it should not be taken for granted that there will be full observability of LV networks. As a consequence, the methods for phase allocation proposed in this project are chosen on the grounds that this situation is not going to change in the short term, and how accurate these approaches can be despite this.

3 Phase Identification Problem

This section addresses the phase identification concept and the phase unbalance of LV networks. It is detailed how unbalancing happens in LV networks, as well as its consequence and significance.

To understand the significance of the phase identification problem, distributed systems were firstly introduced in section 2, where it is discussed how they are evolving towards more complex systems where DG and SMs will have an important role to play. As mentioned in this previous chapter, the increasing presence of DG, and particularly of PV-DG, results in phase imbalances. In this chapter the concepts of phase imbalance and phase identification, in addition to their causes and impacts, are explained.

3.1 Unbalanced Networks in the UK

Most domestic consumers in the UK are connected to the LV network via single-phase cables taken from the three-phase cable, and whose voltage tolerance varies from 6 to 10 per cent of the reference voltage 1pu. However, there is usually little or no information about the phase to which these households are connected to. In fact, one of the main challenges for the optimal functioning and management of these networks is missing customer phase information (Mokhtar, et al., 2019).

Although it is usually assumed that these networks, are balanced, according to (Kong, et al., 2018), more than 70% LV network have three-phase unbalance up to some degree. Moreover, as discussed in several studies, the major causes for the unbalancing are originated at the LV side (Ma, et al., 2015) (Ochoa, et al., 2005).

It should be noted, that even if systems were initially balanced, and the phases to which each customer is connected were known, due to all the factors that can cause phase imbalance, these phases may change. In addition, since DNOs no longer have the correct information about the phases, then they cannot do anything to try to balance back the network. And what is more, they may not know the network is unbalanced.

3.1.1 Causes of Unbalanced Systems:

LV networks are commonly unbalanced, even without the introduction of DG, since household appliances connected to single-phase lines can produce the unbalancing. This

may be because over time some loads are removed, while others are added, resulting in unbalanced loads. Sometimes, even during the same day, there can be fluctuations in the phases of the feeder (Kai, et al., 2013)

Other factors such as asymmetrical line configurations (Yan & Saha, 2013), or the already mentioned presence of PV-DG in the LV circuit can also generate phase unbalancing, as shown in several studies such as (Olivier, 2018) (Schwanz, et al., 2017),

3.1.2 Consequences of Imbalances in LV Network:

Three-phase imbalances can lead to more inefficient utilisation of assets (Zhu, et al., 1998). Concretely, in LV feeders the phase with higher power will exhaust the per-phase capacity. Also, it can affect the efficiency of transformers, since imbalances decrease its available capacity. Both of these consequences, will at the same time result in having to strengthen the network more often that it would be necessary for balanced circuits (Ma, et al., 2017) (Ma, et al., 2015).

With regards to the cost of unbalanced systems, the losses caused by the imbalance can be very costly for utilities, and hundreds of millions of pounds are spent to compensate the unbalanced systems. (Kong, et al., 2018) (Ofgem, 2010)

Furthermore, several studies (Strbac, et al., 2014) (Strbac, et al., 2018) indicated the phase imbalance impacts on the level of losses in the network. As a matter of fact, these studies showed that even small imbalances result in power losses, but as depicted in Figure 3.1, the relationship between phase unbalance and losses is not linear, and large imbalances, for example, 25% of imbalance cause 30% of losses.

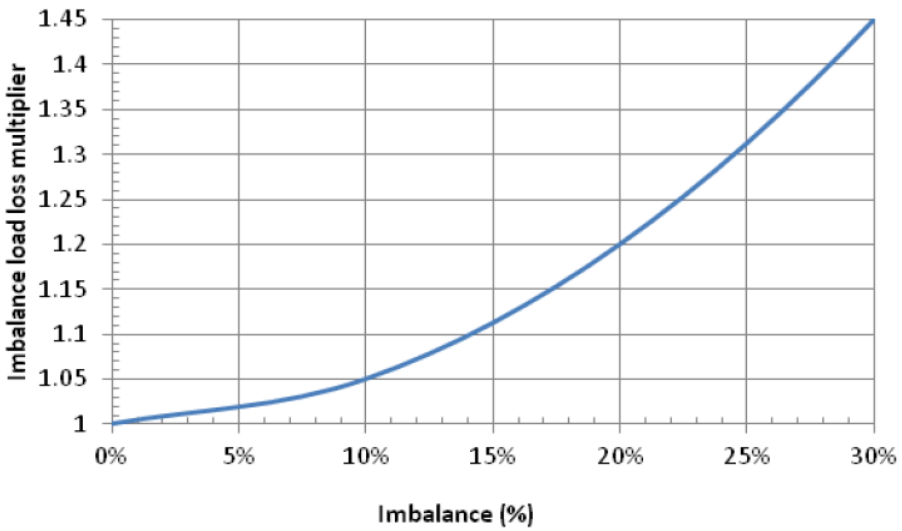

Figure 3.1 - Phase imbalance impact on losses (Strbac, et al., 2014)

On another note, imbalances can reduce the hosting capacity of DG in the network, that is the number of embedded technologies can be deployed without affecting the correct operation of the system (Dubey, et al., 2015), (Cundeva, et al., 2016).

Last but not least, imbalances can also risk the correct functioning and safety (Ochoa, et al., 2005) . Besides that, it can lead to higher operating temperatures, achieving the thermal capacity of the network, which is a situation that would not happen if the circuit was balanced, and also leads to shorter lifespan of the network assets (Ma, et al., 2015).

3.2 Phase Identification and Phase Balancing

In light of the above, it becomes clear that in order to avoid unbalanced systems, phases to which consumers are connected needs to be known. This is where phase identification techniques come to play. If the phase allocation is accurately and precisely done, DNOs could take action to balance the system. Apart from, avoiding the negative effects of unbalanced systems, balance systems can enhance existing facilities and defer feeder expansion projects (Zhu, et al., 1998). Balanced systems are more reliable and therefore, would make DG more attractive to customers, and could lead to better power quality and decrease in the electricity costs.

Regarding how phase imbalance can be handled, the next chapter revolves around different phase identification approaches.

4 Literature Review on Existing Methods

This chapter provides a thorough revision of the existing solutions for solving the phase identification problem concerning their applicability to the study cases. Due to the nature of this project, reviewing phase allocation approaches is a central point for understanding and choosing the best method, and how to adapt it to the given conditions of this project.

4.1 Introduction

There is a considerable amount of existing methods for solving the phase identification problem in literature. These methods may vary significantly, depending on the available data and equipment. For this reason, before going into detail about phase identification methods, it is necessary to provide a summary of the characteristic of the test networks that will be studied, in order to understand which methods could be implemented.

Six case studies representing domestic LV network in the UK will be analysed in this thesis. These networks are discussed in more detail in chapter 6, but here are their main features:

- These networks are composed of customers connected to the same distribution transformer. These LV networks are relatively small, having from 31 to 111 customers. Varying levels of topology complexity are present in the network set, in order to test how this would affect the approaches.

- Discrete half-hourly voltage profiles of customers are available thanks to SMs located at the consumer point. There is full coverage of SM in these networks; that is, there is voltage data from all the customers connected to the transformer. However, lower levels of penetration are intended to be simulated to determine how the phase identification solution would be affected.

- There is no information about the voltage measurements at the transformer point that could be used to correlate both measurements.

- In addition, power measurements both at the transformer and customer point are not available.

For the sake of analyses in this thesis, the phases connection of all SMs is known, in order to be able to determine the accuracy of the phase allocation by the phase identification method to be applied.

Compared to similar phase identification studies, the data available in this thesis is limited, and thus, most of the methods outlined next are not applicable to the test networks for this reason.

Nevertheless, and comprehensive overview of the research realised on the methods for solving the phase identification problem and their suitability for this project are presented below. In addition, note that a summary of the main finding derived from this investigation is presented at the end of the chapter.

4.2 Automatic methods

Existing methods can be divided into two big groups: Manual methods, which require technicians to go to the SM site in person, and automatic methods, that use the available remote information on the network. Also, the latter may rely on additional equipment to acquire more data, which is not usually installed in domestic LV networks, as is discussed further in subsection 4.2.2.

4.2.1 Manual Methods

These methods require the technician to physically go to the premises and read the phase identifiers measurements. These devices are deployed in the point of the interest in the network, to enable the technician to identify the phases to which each household is connected to (Olivier, et al., 2018).

4.2.2 Automatic Methods

Automatic methods can be divided into five subcategories, depending on the type of data used for allocating the phases, as follows:

- **Remote Phase Identifier Technologies**

Following the manual method using phase identifiers, there are new technologies, such as SMs that use PLC (Power Line Carrier). This technology consists of a conductor transmitting the information from the consumer point to the receiver point. In addition, it allows to identify the phase at the consumer end, and process this data (Gunasekaran, et al., 2017)

The main drawback of these technologies is their costs. Also, most of the SMs nowadays, do not include this technology, and issues could be derived from adapting the infrastructure for their installation (Olivier, et al., 2018).

- **Methods that Require Additional Equipment**

The approach put forward by (Wen, et al., 2015) requires the installation of microsynphasors to solve the phase identification problem, since these devices allow to measure voltage phase angles and voltage magnitude. The main reasoning of this method is that if a customer SM is allocated the correct phase, this should have the highest correlation for voltage magnitude and phase between the customer SM and the transformer.

Methods like this one, yet very successful and straightforward, were not considered for this case study, since this equipment is not commonly found in practice, and it would be very costly to deploy it in the current infrastructure. Plus, as already mentioned, in this thesis, test networks represent the current situation in the UK, which do not include this type of devices.

- **Methods Based on Power Measurements**

There is a significant number of phase identification methods that revolve around comparing power measurements from the customer SM and the transformer, or the summation of all the power measurements at the customer side, as it is further detailed in the methods presented below. Although all these methods have proven to be effective in practice, and SM that provide power measurements are common in the market, the majority of the SM which have been deployed in the UK do not provide power measurements. Consequently, the methods discussed in this section are not applicable to the test networks under study in the thesis.

- (Arya, et al., 2011) computes customers phase merely based on time series of power measurements from both the customer SM and the transformer. It is founded on the assumption that the sum of power measurements of customers amounts to the measurements at the transformer. The method involves integer programming and search algorithms, considering both noise-less and noisy scenarios. This is done by injecting a unique signal into the phase line that enables

to identify the phases. Although the results for this method proved to be highly accurate, it is also true that this method is not feasible at the large scale at this moment. This because the type of SM technology presented in this method is not common in practice, and higher cost would be incurred to introduce this technology in the current infrastructure.

- (Fan, et al., 2012) apart from discussing how machine learning and data mining techniques can be applied for the development of smart grids, a novel approach for resolving the phase identification problem based on these techniques is presented. This method is based upon time series of power measurements from the household SMs uniquely. This approach was developed from modifying the method presented in (Arya, et al., 2011). The main difference between both methods is that in (Arya, et al., 2011) they also count with power measurements from the transformer. Meters at the transformer level are not common practice, and therefore this approach presents a successful method for allocating phases for this additional constraint.

- (Chen, et al., 2012) employs voltage phases angles measurements on the secondary side of distribution transformers to determine the connecting phase of the transformers.

- Unlike the methods discussed in this subsection which use time-series measurements, (Brint, et al., 2018) uses half-hourly data of grouped meters are available to network operators. Therefore, this approach is based upon power measurements from the grouped customer SMs and the transformer reading corresponding to the same phase, from a data set composed of 96 users. It was proven to allocate phases with high accuracy to the grouped meters.

- **Methods that Compare Time Series of Voltage Measurements between SMs and Transformer Readings.**

As already discussed, the LV networks under study already only present SM that provide half-hourly voltage measurements from the customers' households. Consequently, although the approaches included in this subsection are the most commonly used, it was not possible to apply them in this thesis.

Hereafter, a wide variety of automatic methods that fall into this category are listed:

- (Seal & McGranaghan, 2011) analyses voltage measurements with the object of finding changes in the voltage magnitudes and be able to identify the phase of the loads by comparing these measurements with SCADA SM data. This method proved to work for both single-phase and three-phase customers, and for a network formed by 75 customers.

- (Pezeshki & Wolfs, 2012) compares time series of voltage measurements from a three-phase household SM with measurements from other single-phase customer SMs on the same network. Correlation is applied to both time voltage profiles from a signal processing outlook. This method allows to obtain the phase to which the 3-phase SM is most likely connected to. No additional equipment is required, and it was proved successful for a LV network with 75 consumers.

- In (Pezeshki & Wolfs, 2012) method, time-series voltage measurements from both the distribution transformer and the customer SMs are analysed with correlation method, in order to allocate phases of the household phases. It is based on the assumption that voltage profiles of customers are well correlated with the voltage profile at the transformer level on the same phase. Note that these assumptions were proven to be right both in the simulation and in-field results.

- While most automation methods rely on the availability of precise connectivity models, which cannot be always accessible, largely due to changes in the network. (Arya, et al., 2014) expands previous work, and in this method infer phase connectivity from voltage measurements of the customer SM only, or in addition to measurements from the medium- voltage network, and partly accurate information about the phases allocated to each customer. This method provides a 93% of accuracy when the measurements set is large enough (i.e. couple of days).

- **Clustering Techniques**

On another note, there are a number of methods which use clustering techniques based on correlation coefficients. The methods proposed in this thesis belong to this type of phase identification approach. Therefore, a more detailed description of this machine learning techniques to be used in the analyses is provided in section 5. Hereafter, an overview of already successfully tested methods is provided:

- In (Arya & Mitra, 2013) a novel method using k-means clustering is proposed, where time series of voltage measurements from customer's SMs are used for identifying the phases of each customer. In addition, the k-means algorithm is also used to create sub-clusters for those customers that belong to the same building and are connected to the same phase, in order to find connectivity relationships within the LV circuit. A similar approach is proposed in section 5, since k-means clustering is one of the preliminary methods. However, the data set to which the algorithm is applied is different since it should be recalled, that data to be used is composed of half-hourly voltage.

- (Short, 2013) approach applies linear regression on voltage SM and substation measurements. This method is based on correlating the aggregation of SM measurements with substation voltage observations, and is topology-sensitive, which can lead to inaccuracies. Although this research sheds light on the relationships between the measurements at the household and substation level; once again, this method is not suitable for this thesis, since required data to perform it is missing.

- The results obtained with this approach, (Mitra, et al., 2015), demonstrate that time-series voltage SM data is correlated with the hierarchical structure of LV networks. This correlation is used with the k-means algorithm to partition the measurements into three groups and enable to identify the phases of the SMs. This method is the most similar to one presented in this thesis. The main difference lies largely in the type of data analysed, since even though this approach only uses voltage measurements from customers, these are time-series voltage profiles,

- As already mentioned, connectivity models can become outdated due to changes in the network topology. Consequently, (Bandyopadhyay, et al., 2015) proposes a new method for inferring customer phases from voltage measurements of customer SMs using machine learning techniques. In this case, supervised, semi-supervised and unsupervised techniques are tested on real SM measurements from a LV network in North America.

- In (Wang, et al., 2016) k-means clustering is successfully deployed, based upon customer voltage measurements. Instead of directly employing correlation, it extracts features on which they perform a k-means clustering, achieving to divide

customer SMs into clusters corresponding to the different phases with an accuracy of over 90%.

- (Watson, et al., 2016) presents a promising method that enables to use voltage measurements from customer SM to determine the transformer, as well as, the phase to which each customer is connected to. An algorithm that uses correlation analysis with the Fisher Z transform is presented to achieve this.

- (Jayadev P, et al., 2016) puts forward a novel data-driven approach that employs graph theory and PCA to identify phases of a simulated set, considering the effect of noise. From these noisy measurements, the algorithm has been proven to efficiently identify the phases of the household SM.

- (Olivier, et al., 2017) employs graph theory, in this case, maximum spanning tree technique, in order to identify the phases of SMs in a LV network, using uniquely the time-series voltage measurements. In this way, since distribution networks are usually shaped in radial fashion, the branches of the resulting tree represent the cable lines, and the nodes are the buses. The algorithm is used to maximise the correlation coefficient between the voltages at the nodes, and based on the results obtained, group the measurements into 3 clusters.

- (Pappu, et al., 2018) method relies on energy measurements to identify the underlying network topology, so that phase of SM can be allocated correctly. It applies graph theory based on principal component analysis (PCA) to SM energy measurements.

- (Olivier, et al., 2018) method expands the work developed in (Olivier, et al., 2017). Voltage measurements are analysed using a clustering method based on graph theory and k-means algorithm but modified in order to take into account the underlying structure of LV network, even when its topology is unknow. It also compares the results obtained from this novel approach to those with the k-means method used in other studies. Obtaining a higher accuracy with their proposed method for different scenarios, such as, when there is a greater presence of 3-phase SMs than single-phase SMs in the circuit, or when the average voltage profiles are calculated for periods lower than one minute.

- (Blakely, et al., 2019) puts forward a machine learning, spectral clustering method, for identifying and correcting phases allocated to customer SMs. Spectral clustering is employed in combination with a 'sliding window ensemble' which improves the accuracy, and its suitability for larger networks.

Given that many of the above-mentioned methods used k-means and graph theory have been proven to be successfully implemented in voltage measurements only, it was concluded that these methods could be suitable for discrete voltage measurements. Based on the fact that it has been proven that SM connected to the same phase present high values of correlation.

4.3 Findings from Literature

First of all, it should be highlighted that although some of the methods presented in this chapter would be outstanding options for solving the phase identification problem, they were considered to be not suitable for the case studies of this thesis.

For example, installing SMs which include PLC technologies would be a great option to sort out the phase identification problem, and have a higher level of monitoring in LV networks. However, it is not within this project's scope to determine which technologies would be more beneficial to include in LV networks for their better management and operability.

What is more, the test networks in this thesis represent the real situation in the UK nowadays, and it is of great interest to research if it is possible to apply a method that does not require any additional change in the SM infrastructure. This is because even if these changes in the infrastructure would be found to be cost-effective and beneficial, they are not likely to happen any time soon. And it is worth recalling that the objective of this thesis is to present a phase identification solution that can work for the current situation.

On top of that, many methods used time-series based approaches, and given the fact that only discrete half-hourly measurements from the customers are available, these methods had to be discarded as well.

Therefore, the only automatic methods that were found to be applicable to the case study were those that rely on clustering techniques for identifying the phases of SMs.

In addition, since only data about voltage profiles from customers is available, it was found that the only criterion that could be applied for the clustering of the data set is based on using correlation factors. This is because it has been proven that voltage profiles of customers connected to the same distribution transformer present high correlation values. In addition, it was found that customers connected to the same phase are expected to have the highest correlation (Luan, et al., 2015).

All in all, it was concluded that the only suitable techniques are clustering approaches based on grouping the SM voltage profiles according to their correlation factor. Meaning that the clusters to be formed will be composed with the SM with higher correlation factors among them, and therefore, the SMs belonging to the same cluster would be connected to the same phase.

In particular, hierarchical clustering and k-means clustering are proposed to be applied in the test network. Both methods have been successfully used in a considerable number of studies, especially the k-means algorithm which a very commonly clustering technique. In addition, they can be easily programmed to used correlation criterion for the clustering, thanks to software like Matlab which enables to use machine learning techniques.

Although similar approached worked in studies like (Mitra, et al., 2015), which uses correlation-based k-means algorithm; it is yet unclear up to what point and under which conditions, these correlation values would be enough to form clear clusters for a more data-constrained scenario, such as the ones presented in this thesis.

This is tested in section 8, 9 and 10, where the proposed methods are applied to the case studies and their capability of identifying SM phases correctly is investigated for several scenarios, of particular interest is how PV-DG can affect the accuracy of the methods. In addition, in the next chapter, the rationale behind these two preliminary methods is further detailed.

5 Description of the Proposed Methods

This chapter outlines the preliminary machine learning techniques chosen to tackle the phase identification problem, that is, hierarchical and k-means clustering. Besides, the techniques used to evaluate the quality of the cluster used in Analysis 1, 2, and 3 are also explained. In addition, the improved k-means algorithm employed in Analysis 3 is described as well.

5.1 Introduction

After reviewing methods from literature, it was concluded that considering the constraints and type of data, these unsupervised machine learning techniques could be proved to work for solving the phase identification problem. It should also be noted that the two methods presented below, would be tested in practice with the objective of determining which of them provides more accurate results for matching results.

These two machine learning methods fall into the category of unsupervised techniques, which are described to understand the suitability of this type of methods for the case studies of this thesis.

5.2 Unsupervised Machine Learning Description

Unsupervised learning techniques are a type of machine learning approaches which allow identifying the natural patterns or groupings in a data set. In this case, the voltage measurements from the customers SM need to be grouped in 3 different clusters according to the phase they are connected to, uniquely based on the patterns that will be identified by analysing their voltage profiles based on correlation factors.

The main goal of clustering is to disclose hidden structure in a data set (Vathy-Fogarassy & Abonyi, 2013). Clustering algorithms use this structure to divide the given data into groups, so that objects belonging to the same group or cluster are the most similar to the other objects in the same group, and dissimilar to those in other clusters (Berkhin, 2006).

These techniques can be used, in addition to finding intrinsic structures in data, as a method for initialising regression models, classification of data, correlation analysis, and anomaly detection among other utilities (Vathy-Fogarassy & Abonyi, 2013) (Han, 2006).

Clustering approaches may be divided into two main groups: hierarchical and partitioning.

(Jain, et al., 1999). The former generates different levels of partitions, placing at the top of the sequence, only one cluster including all the objects of the data set, and at the bottom, all the clusters composed by one object. Furthermore, hierarchical techniques can be split into agglomerative and divisive as well.

Partitioning, however, strives to divide the data set into non-overlapping clusters, so that the dissimilarity between clusters is as high as. The most commonly used partitioning method is K-means clustering (Cui & Potok , 2005).

Both agglomerative hierarchical and k-means clustering techniques are tested for the case studies of this project and are explained in detail below.

5.3 Hierarchical clustering

Hierarchical clustering partitions data sets into groups, and those groups into larger groups, resulting in a hierarchical nesting of clusters. Dendrograms are the illustrations representing the hierarchical tree where the branches are the observation belonging to the data set, and the nodes are the clusters. This method allows to investigate the groupings that can be obtained for different granularities.

As previously mentioned, hierarchical methods are further divided into agglomerative and divisive. The difference between these techniques boils down to the structure and operation of the algorithm. Divisive (top-down) approaches starts from a single cluster enclosing all the observations and proceeds to gradually partition the initial cluster until the stopping criterion is met. On the contrary, divisive (bottom-up) clustering starts from cluster formed by a single observation, and then combines clusters with the highest similarities to create bigger clusters until the stopping criterion is met too. It should be noted that the stopping criterion is usually achieved when the number of k clusters is met (Berkhin, 2006). It should be noted that the agglomerative approach is the one applied in this thesis, and deployed using Matlab functionalities, as explained further below.

In Figure 5.1, it can be observed the results that can be obtained using hierarchical algorithms. At the left of the Figure, the observations and their assigned cluster are represented, while at the right, it is shown the resulting dendrogram tree which represents the nested clustering of the observations.

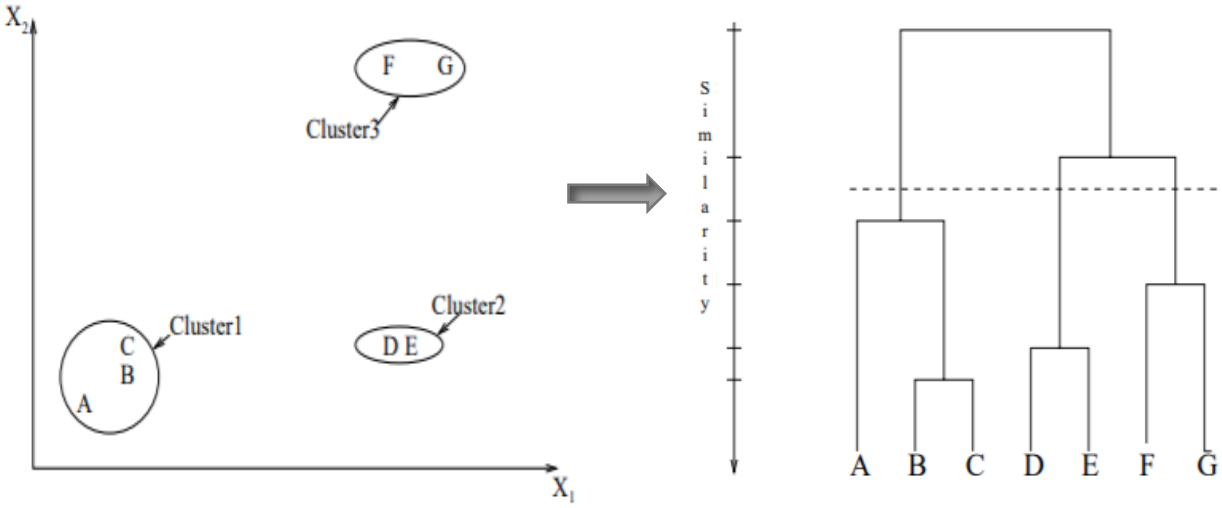

Figure 5.1 - Clusters and the dendrogram obtained using the hierarchical algorithm (Jain, et al., 1999)

5.3.1 Hierarchical Algorithm in Matlab

Here is explained, how the agglomerative hierarchical technique works and can be implemented in Matlab, since this is the way it was employed to perform the analyses in this thesis.

As above mentioned, the clustering algorithm splits data into a hierarchical multilevel tree or dendrogram, which enables to find the most suitable level of clustering for different conditions (MathWorks, 2018).

By installing the Statistics and Machine Learning Toolbox in Matlab, hierarchical clustering can be applied to a given data set. The following Matlab function enables to apply this clustering technique, by following the next steps:

1. Firstly, by using the ‘pdist’ function to calculate the distance between the points. In other words, in order to later cluster the observation, firstly it is needed to compute how close each pair of observation are to others. This calculated by using correlation distances, although other methods such as quadratic Euclidean distances could be used. For this thesis, however, the ‘pdist’ function is programmed to use the correlation coefficient to measure the similarity between observations, as it can be observed in Figure 5.2,.

2. Secondly, the 'linkage' function is used to start the clustering by initially forming clusters formed by only two pair of observations: This pairing is based on similarity degree computed for each pair of observation in the previous steps. This process is then continued by merging cluster whose observations are closest among them, and this is repeated until the forming the hierarchical tree.

3. In order to determine the optimal clustering, the function 'cluster' is used to prune branches from the bottom of the tree up to the clustering level specified as an input to the function. For this project, the objective is dividing the data set into three groupings, and this is specified in the Matlab code, as it can be seen in Figure 5.2.

```
% The linkage function calculates the distance between each pair of points and, using these distances,
% determines the tree hierarchy by linking together pairs of "neighboring" points
Z = linkage(VoltageArray');

%dendogram allows to  visualize the hierarchy of a binary cluster tree.

dendrogram (Z);
%cluster function is used to assign observations into groups according to the linkage distances Z
group1 = cluster (Z, "maxclust",3);

Y = pdist(VoltageArray', "correlation");

%The Cophenetic correlation coefficient quantifies how accurately the tree represents the distances between observations.
% Values close to 1 indicate a high-quality solution.

c = cophenet(Z,Y)
```

Figure 5.2 - Example Matlab code for hierarchical clustering

5.4 K-means Clustering

As already mentioned in section 4, k-means clustering is being very successful in allocating phases for LV networks.

5.4.1 Description of k-means Algorithm

K-means clustering (MacQueen, 1967) is a commonly used approach when the number of groups to be formed is known. Since this algorithm allows to divide an M-dimensional data set into a predefined k-number of groups (Wagstaff, et al., 2001). This approach is very suited for phase identification problem since it is known that there can be three different phases connections, and therefore, the partitioning of the data set into three clusters is sought.

This is done by firstly randomly determining the centre of the clusters. There are as many centroids as specified clusters. Then, each observation in the data set is assigned to the closest cluster centre. This means that the k-means method is sensitive to the initial location of the centroids, since the algorithm finds the optimal solution for this initial location.

For this reason, the algorithm needs to be iterated, using different initial cluster locations until there is so as to the centroids converge to the same location after a number of iterations or the squared error decrease is negligible, that is the optimal clustering is achieves, as illustrated in Figure 5.3 (Berkhin, 2006)

.

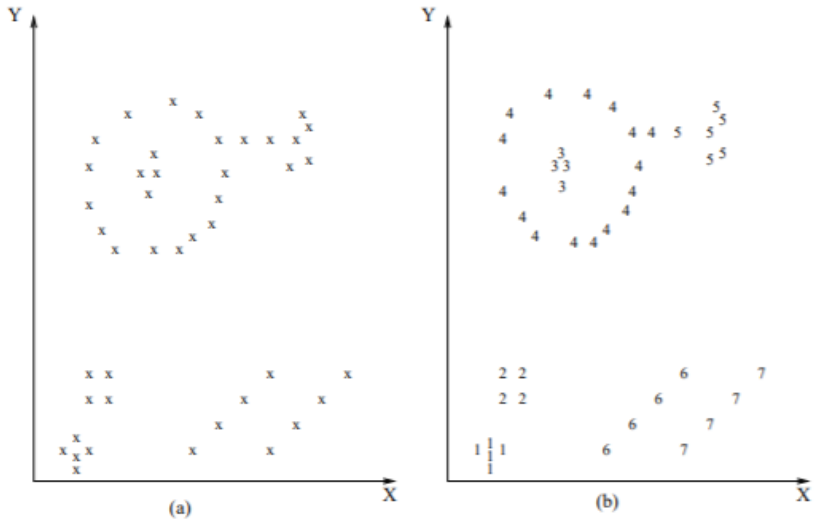

Figure 5.3- Example of k-means clustering algorithm (Jain, et al., 1999)

Equation (5.1) (Vathy-Fogarassy & Abonyi, 2013) represents how this algorithm works, where:

- C_i, is the cluster number ith.
- $\|x_k - v_i\|$, is the selected measurement for distances between the observation (x_k) and the cluster centroid (v_i).

$$J(X,V) = \sum_{i=1}^{c} \sum_{x_k \in C_i} \|x_k - v_i\|^2 \qquad (5.1)$$

The k-means algorithm is very simple to implement since it is easily processed by digital computers. As a matter of fact, k-means is even embedded in the Machine Learning Toolbox for Matlab, which was used to develop this project.

5.4.2 K-means Clustering in Matlab

The k-means algorithm can be employed in Matlab (MathWorks, 2019) as well, by using the function 'kmeans', which likewise the mathematical algorithm, groups the data into k clusters. Yet, this function also yields a vector as an output which relates each observation with its corresponding cluster index. The code in Matlab looks as depicted in Figure 5.4:

```
%k-means function

grp2_1 = kmeans (VoltageArray2', 3, 'Replicates',200, 'Distance','correlation');
scatter3(scrs2(:,1),scrs2(:,2),scrs2(:,3),10,grp2_1)
view(164,16);
```

Figure 5.4 - Example of code in Matlab for k-means clustering

While hierarchical clustering operates with multilevel dissimilarity distances, k-means works with the actual data-set objects, resulting in single-level groupings. Consequently, k-means method tends to be more adequate for bigger data sets.

The 'kmeans' function algorithm in Matlab *"uses a two-phase iterative algorithm to minimise the sum of point-to-centroid distances, summed over all k clusters"* (MathWorks, 2018).

- 'Batch updates' is the first phase algorithm, which assigns all the observations to the closest cluster centre simultaneously in each iteration, and then calculates the new centroids. It may happen that the 'batch updates' phase does not meet the local minimum, i.e. when any observation cannot be moved to any other cluster where the sum of distances would be higher. This is sometimes the case of a small number of observations. Although this first phase does not usually yield the definite solution, it is used to get a fast approximation that will be considered in the second phase algorithm.

- In the second phase, the 'online updates' algorithm is used. This method reassigns the cluster index of each observation, only when it would decrease the sum of distances. If so, the centroids are recalculated as well. Unlike the first phase, the 'online updates' does converge to the local minimum. However, this does not mean that there could be another local minimum which is more optimal.

As above-mentioned the solutions obtained with the kmeans function is dependent on the initial cluster centres. In Matlab, to avoid the local minimum, it can be specified the

number of iteration, by using the 'Replicates' option (also used in the code shown in Figure 5.4). In addition to this, in the kmeans function it can also be specified the type of distance measures. If not specified, squared Euclidean distance metrics are computed. However, to analyse voltage profile correlation, measures are more suitable, since voltage SM data connected to the same phase have higher correlation values, when compared to voltage profiles corresponding to other phases.

5.5 Methods for Evaluating the Clusters

The main objective of cluster analysis is to determine the optimal number of clusters. Therefore, these techniques can very helpful tools for unsupervised techniques. Since the main drawback of k-means and hierarchical clustering is the fact that the number of clusters needs to be defined.

As it can be seen in section 8, Analysis 1 uses both hierarchical and k-means clustering in one of the test networks with the goal of comparing their resulting phase identification outcomes. Within analysis 1, two cluster evaluation techniques are applied, silhouette and confusion matrix, which are detailed below.

5.5.1 Clustering Evaluation Techniques Used in Analysis 1

Silhouette Plot

The silhouette plot (MathWorks, 2019) represents in the x-axis the silhouette value, between -1 and +1, which is used to compare how close each observation to other observation in its same cluster, with observations in the other clusters. In this case, the function was programmed to calculate these distances based upon correlation approach.

Silhouette plots are used to analyse the quality of the cluster in Analysis 1, as tools for comparing the results obtained with k-means and hierarchical clustering.

Confusion Matrix

In addition, although confusion matrix was used for the first analysis, with the goal of representing the accuracy of the phase matching for both methods.

Confusion matrixes can be generated in Matlab using the function 'plotconfusion' (MathWorks, 2019). This function enables to compare the target outputs of the algorithm with the real results. Therefore, a matrix with the real known phases corresponding to

each SM, as well as, a matrix with the phases allocated with the hierarchical algorithm will be compared.

The columns in Figure 5.5, correspond to the target class (results), while the rows correspond to the estimated outputs with the hierarchical clustering. Also, while the diagonal cells are those observations that were correctly allocated their phases, the other cells correspond to the amount of SMs that were allocated the incorrect phases.

Furthermore, the column on the right in addition to the row at the bottom, show the successful (green percentage) and failure (red percentage) rate corresponding to each class.

Figure 5.5 – Example of confusion plot

5.5.2 Clustering Evaluation Technique used in Analysis 2 and 3

After comparing the results yielded both by hierarchical and k-means clustering, as discussed in section 8, it was decided to proceed with the next analysis with just the k-means clustering.

In Analysis 2 and 3, therefore, the accuracy of the phases allocated with the k-means algorithm are analysed for different factors. In order to represent the varying accuracy levels, bar chart and radar charts are used. It is of particular interest to explain the use of radar charts.

Radar charts (also called spider plots) allow to illustrate how more than 3 variables affect each other. In these analyses, they were used to depict how the percentage of each of the three phases, as well as, the total percentage of SM coverage in the network, affects the accuracy of the phase identification.

5.5.3 Improved Algorithm Description

For reason later explained in the results chapter, the originally proposed algorithm was decided to be improved. This is because sometimes, for lower levels of SM penetration it may happen that the available SM are only connected to two different phases (or even one). Since k-means algorithm needs to be inputted a defined number of clusters to be formed, this is a huge drawback when only two phases are required to be identified.

Therefore, as a solution to this issue, it is proposed to use the gap clustering evaluation criterion. As explained below, it allows to determine the optimal number of clusters.

Gap Statistic Criterion.

Gap statistic, as defined in (MathWorks, 2018)is based on the graphical method of representing error measurements versus the number of potential clusters, with the objective of finding the 'elbow' between both variables. This point occurs when the highest gap value for a number of clusters is achieved. This gap value is defined in equation 5.2:

In addition, the function 'evalclusters' in Matlab can be programmed to use gap criterion, as well as the potential number of k-clusters to evaluated for a given data set. This function creates an object, which includes the variable OptimalK, which can be used as input for the k-mean function. This way, the k-means algorithm will always partition the data into the optimal number of groups. The code required to do so is shown in Figure 5.6:

```
evaluationGap = evalclusters(VoltageProfileM','kmeans','gap', 'Distance','correlation', 'KList',[1:3]);
Accgrp = kmeans (VoltageProfileM', evaluationGap.OptimalK, 'Distance','correlation','Replicates',100);
```

Figure 5.6 - Evaluation Gap code in Matlab

Therefore, this is the improvement proposed for the k-means algorithm, which is supposed to yield better results when the number of different phases to be identified for a SM set is lower than three. This improved algorithm will be tested in practice in chapter 10.

6 Description of the LV Test Networks

This chapter describes the LV distribution networks that have been used to analyse the proposed methods. It is detailed its topology, the type of available data, and how this data was generated.

In this thesis, six different LV networks will be used as case studies for the proposed methods. These networks are existing LV circuits selected from Scotland Central Belt, although the voltage measurements used for the methods, were synthetically obtained and provided by Scottish Power, as explained in 6.2.

In this chapter, the topologies of the six circuits are firstly described, followed by an explanation of the type of data available for the test circuits; and finally, the method employed to obtain the voltage profile is presented.

6.1 Main Characteristics of the LV networks

These six circuits were selected as case studies with the objective of representing different levels of topology complexity, that is varying degrees of branching and the number of households connected to the network. Furthermore, the presence of PVs installed on the rooftops of these households is represented in three of the test circuits.

Although already mentioned throughout the thesis, it is worth stressing, that these circuits were chosen to represent the current data-constrained situation in the UK for network operators. Where:

- Only half-hourly voltage profiles from the customer's points in the network are available. This is because the current SM settings are programmed to send information to the network operators each half an hour, and also because of privacy-related issues.

- These circuits are lacking measurements at the transformer point, since apparently this a widespread situation even nowadays.

- Note that the circuits represented are formed by domestic users uniquely, and all the SMs represented in the circuit are connected to the same transformer, and the voltage profiles for all customer’s SM are available, as if there was a 100% level

of penetration, which will allow to simulate varying levels of penetration in chapters 8, 9 and 10.

- It should also be noted that the type of voltage measurements for the three first circuits is instantaneous voltage, while for the last three networks is the half-hourly average voltage. Note that the current setting of SM establishes that the readings sent to network operators correspond to the average measurement. Although this could be easily changed if better results are obtained for a type of data.

The following are the circuit's topologies and their specific characteristics:

LV Circuit 1

Figure 6.1- Topology of the LV Circuit 1

Circuit 1 is the least complex LV network, composed of 31 SMs[3] with a limited level of branching. The length of the feeder is 450 m.

There is no presence of PV-DG in this circuit. In addition, instantaneous half-hourly voltage measurements are generated for this circuit.

[3] Note that the numbers in the LV circuits represent the number of SM per building, that is, the number of customers connected to the network in that point.

LV Circuit 2

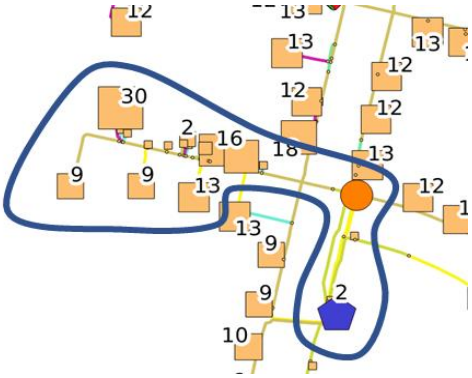

Figure 6.2 - Topology of the LV circuit 2

The length of the feeder of circuit 2 is approximately 300 m, and supplies 56 customers

Like circuit 1, there is no presence of PVs on the rooftops, and half-hourly instantaneous voltage data is available.

This circuit also presents a low level of branching and therefore, limited topology complexity.

LV Circuit 3

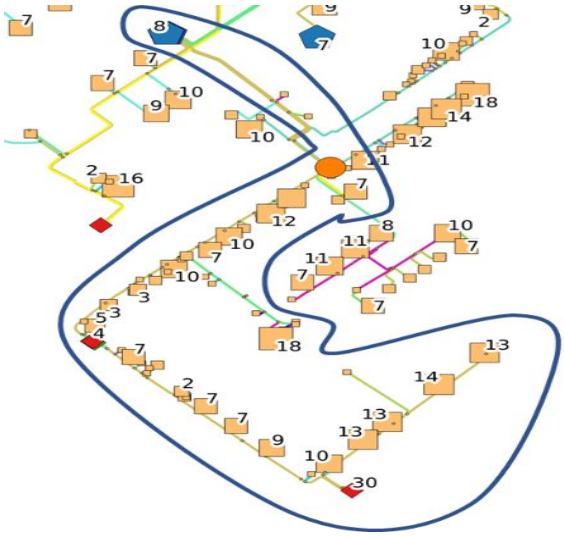

Figure 6.3 - Topology of the LV circuit 3

The length of the feeder of circuit 3 is approximately 450 m, and supplies 56 customers

Likewise, there is no presence of PVs on the rooftops, and half-hourly instantaneous voltage data was generated for this circuit.

The branching in the circuit is also limited, and topology can be considered rather simple.

LV Circuit 4

Figure 6.4 - Topology of the LV circuit 4

Circuit 4 is composed of 111 households, with a feeder length of around 400 m.

This circuit presents a higher level of branching compared to the three previous case studies.

For this circuit the following voltage profiles are generated:

- Half hourly average data for both scenarios with no PV penetration and 50% of penetration.

LV Circuit 5

Figure 6.5 - Topology of LV circuit 5

The length of feeder is of around 560 m and supplies 69 customers.

This circuit also presents a higher level of branching compared to the three first case studies.

For this circuit the following voltage profiles are generated:

- Half hourly average data for both scenarios with no PV penetration and 50% of penetration.

LV Circuit 6

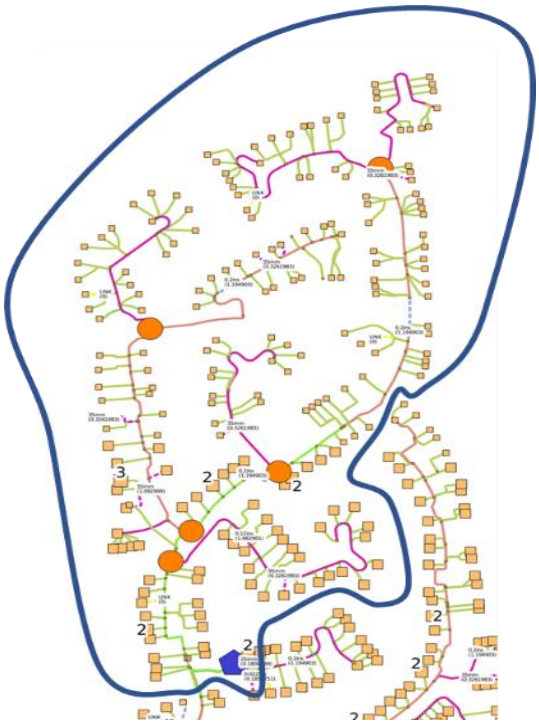

Figure 6.6 - Topology of the LV network 6

This circuit represents the network with the highest level of branching and complexity, among the case studies.

The length of the feeder is approximately 320 m, and it is formed by 63 customer SMs.

For this circuit the following voltage profiles are generated:

- Half hourly average data for both scenarios with no PV penetration and 50% of penetration.

6.2 Generated SM Voltage Data

The above presented LV circuits were randomly chosen from Scotland Central Belt. The process followed to generate the voltage profiles per household for a month is explained in great depth in the article by (Mokhtar, et al., 2019).

Considering the already mentioned fact that the SM coverage in the UK is rather low, this method enabled to generate realistic voltage profiles of customers, using Deep Learning Neural Network (DLNN). This method allowed to obtain the generated voltage profiles for all the customers in the networks, even for those households that did not have SMs, and without needing high granularity demand power data from all customers on the circuit.

In the research presented in (Mokhtar, et al., 2019), the approach was developed by using OpenDSS software, to generate 1-month power measurements without DG.

However, the generated data, provided by Scottish Power, in this thesis although obtained using the same methodology as in (Mokhtar, et al., 2019), was modified for the interest of analysing a more realistic situation for network operators in the UK nowadays. That is, instead of power demand data, the profiles generated are the voltage profiles for a time

period of a month. Depending on the circuit instantaneous or average half-hourly profiles were created since currently SM settings can provide this type of data. Yet, as already mentioned currently SMs in the UK are programmed for measuring average voltage measurements.

In addition to the voltage profiles, there is also information about the real phases to which each SM is connected to, and the distances from the SM to the transformer. The former will only be used in order to compare this real phase with the results obtained with the proposed methods. Conversely, the latter is not really for the development of the algorithm applied to the voltage data, but it could be used to analyse how topology affects the results of the proposed methods, as it is discussed in chapter 13.

6.3 Data Limitations and Assumptions

As mentioned in previous chapters, LV networks traditionally have had little monitoring, and even nowadays there is an existing lack of data that, in some cases, does not enable to apply many phase identification methods presented in the literature review.

The generated voltage profiles that will be used for the analyses in this project are therefore highly constrained and limited. However, it should be highlighted that this is precisely why they are of great interest, since they represent the current situation in terms of data availability for network operators in the UK. And it is very likely that in the near future this situation is not going to undergo significant changes.

6.3.1 Data Constraints

The main data constraints in comparison with similar existing phase identification studies are listed next:

- Lack of measurement reading at the transformer/feeder point
- There is no power demand data from the SM.
- Half-hourly discrete data are only available, which reduces the number of possible methods to apply significantly. This is because many studies use continuous time-series approaches to solve the phase identification problem.

6.3.2 Assumptions

- It is assumed that the measurements from the SMs are synchronised

- All the SMs used in the simulations are assumed to be single-phase. These SMs are the most common meters in domestic LV network at the moment.

- Also, it is assumed that the information of the distribution transformer connectivity is known by the operator network. That is, it is known to which distribution transformer each SM is connected to. All the SMs studied in the six circuits are assumed to be connected to the same transformer.

- The noise in the voltage measurements is assumed to be negligible.

7 Method Description

The following section details the process followed to find the approaches that can effectively identify phases for SM in the test networks. In addition, it details the analysis to be performed in order to assess how different scenarios and variables affect the accuracy of the proposed methods.

As explained in the previous chapter, given the limitations of the available data, most of the existing phase identification approaches are not applicable to the dataset under study.

This is precisely the main challenge, and at the same time contribution of this project, since the methods developed will be adapted to work even in highly-data constrained scenarios.

Consequently, the first step followed, as shown in Figure 7.1, was to carry out a profound revision of all the existing methods, of which an overview is provided in section 4. Next, and considering and understanding the constraints of the given data, two preliminary methods (hierarchical and k-means clustering) were chosen.

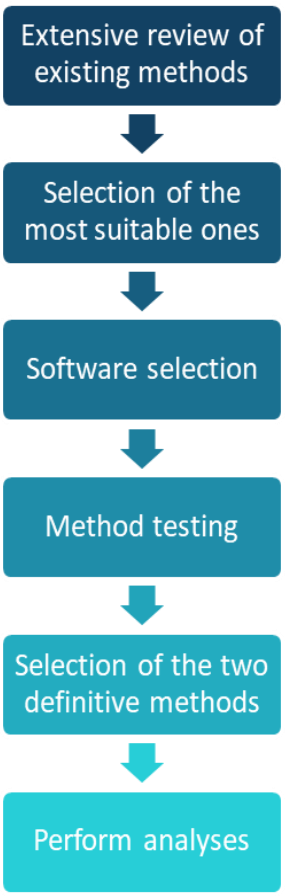

Figure 7.1 - Methodology process flow diagram

It should be noted that the chosen software to implement the potential methods and used to treat the data (which was provided as .csv files) was Matlab. This decision was taken on the grounds that after reviewing literature, the potentially most suitable methods were all machine learning techniques. And considering that Matlab offers a Machine Learning Toolbox, which makes these algorithms easy to implement and adapt to the given conditions and dataset. Other software and tools, such as Python or R, were considered, yet.

Once the software was chosen, the selected methods, k-means clustering, and Hierarchical clustering were implemented in Matlab to test their suitability for the test network. And the next analyses were performed with the objectives of firstly, determine the most suitable method for the cases studied, and then, investigate under what conditions of SM and PV-DG penetration the proposed method would be reliable.

7.1 Analyses Performed on the Test Networks

7.1.1 Analysis 1: Comparison between Hierarchical and K-means Clustering

In order to check which method is more suitable for the given data constrained test networks.

In this case, only LV circuit 1 data set is used for the analyses, so that, the algorithm that is proven to be more effective will be later applied for the rest of circuits and analyses.

Firstly, the algorithms were tested for a scenario of full coverage of SMs in the network. To check if significant differences between the effectiveness of both algorithms could be found. Confusion matrix and silhouette plots were generated in order to visualise how these two methods compared in cluster quality.

On second place, it was tested how lower SM penetration in circuit 1 would affect the algorithms' accuracy. The decrease in the degree of penetration of SM in the network was simulated by reducing the number of SMs in each loop. This was done using two different methods:

- Method 1 consists of randomly knocking out SMs.

  Both clustering techniques are initially applied to all voltage profiles from the customer's SM. But in each loop, one voltage profile (corresponding to one SM) is removed from the data set, until there are no more SMs. This way, it is very

likely that for most of the loops, SMs connected to three phases are present in the data set.

- Conversely, Method 2 knocks out SMs in order, that is, voltage profiles of SMs connected to phase 3 are firstly removed, followed by those connected to phase 2, and finally phase 1. This method allows checking how clustering techniques can work when there is no information about if the SMs to be analysed can be connected to three (or less) different phases.

It is worth noting that the voltage measurements used for this analysis, were half-hourly instantaneous voltage. Also, in each loop, the accuracy of the phase matching was calculated in order to observe how it changes as the penetration decreases.

7.1.2 Analysis 2: Effect of a 50 % Level of Penetration of PV-DG (Method 1)

In this analysis, voltage profiles generated for networks with a 50% level of penetration of PVs are tested. This analysis is very important since one of the main focus of this project is to find successful phase identification methods so that PV-DG could flourish. Since, as previously discussed in section 3, PV-DG is one of the major factors causing phase imbalances, and therefore reducing the hosting capacity of these technologies.

For the above reasons, Analysis 2 is focused on investigating the effect of PV-DG on the accuracy of the phase matching obtained with k-means clusters. In this section, this is only analysed following Method 1, which outputted really accurate results in the first analysis without PV presence.

The average half-hourly SM voltage measurements of LV circuits 4, 5, and 6 are used for this analysis.

7.1.3 Analysis 3: Introduction of Improved K-means Algorithm for both scenarios without and with 50% level of PV penetration.

In light of the results obtained in Analysis 1 and 2, k-means is the chosen method to be used in the remaining analyses. Nonetheless, an improvement to the initial k-means algorithm tested is presented in this chapter.

This improvement consists of using the evaluation gap function from Matlab, which as explained in chapter 6, allows to determine the optimal number of clusters for a data set based on the gap criterion. Therefore, this function is applied to the data set as the number of SMs is decreased in order to simulate different levels of SM penetration. Next, this

optimal number of clusters is used as an input for the k-means algorithm, instead of determining that 3 clusters are to be formed (as it was programmed for Analysis 1)

The potential improvements in the accuracy of the phase allocation by k-means clustering is tested for the six circuits. It is worth recalling that for circuits 1, 2 and 3, half-hourly voltage profiles are used; while for circuits 4, 5 and 6 the instantaneous voltage measurements are instantaneous. Therefore, this analysis will also enable to test how these two types of voltage measurements work for the improved algorithm.

In addition, for the last three circuits both scenarios without PVs and with 50 per cent of penetration will be tested. And the method used for simulating the decreasing levels of SM penetration will be Method 2, since it allows to check how the algorithm will work when the amount of different phase connection is lower than 3.

8 Results for Analysis 1 and Discussion

This section presents the results obtained for Analysis 1, which it is intended to compare the accuracy of hierarchical and k-means clustering. Firstly, both algorithms are compared for full coverage of SMs, and then how their efficiency is affected as lower levels of penetration of SMs are simulated.

8.1 Results for a 100 per cent Level of SM Penetration

The following are the results obtained with hierarchical clustering and k-means clustering for full coverage of SMs in the LV circuit 1:

8.1.1 Hierarchical Clustering

In order to assess the accuracy of the groupings formed with the hierarchical algorithm, several techniques were employed. These are the dendrogram tree, confusion matrix, and silhouette plot, which are presented next:

As explained in section 5, agglomerative hierarchical clustering algorithm can be implemented with Matlab, and enables to group the SM data set into three groups. In addition to this, the algorithm allows for generating a dendrogram, which is depicted in Figure 8.1.

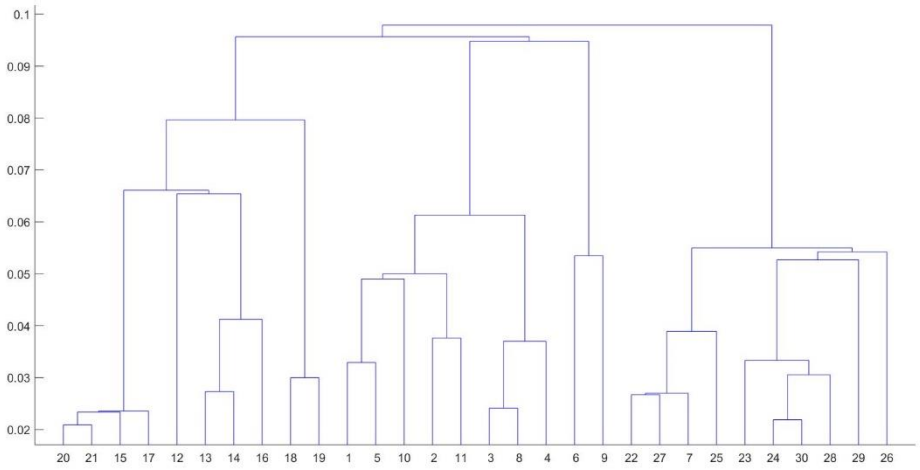

Figure 8.1 - Dendrogram tree for circuit 1 with full SM coverage

Moreover, Figure 8.2 represents the confusion plot outcomes for hierarchical clustering.

This plot was generated to easily assess the accuracy of the algorithms for the different scenarios. As it can be observed in Figure 8.2, the success rate for all the SMs and phases is 100% with the hierarchical algorithm.

Since all the SMs haven allocated the correct phase, the number inside the diagonal phase accounts for the number of SM phases connected to each phase, and the percentage corresponds to the total percentage of presence of each phase in the network.

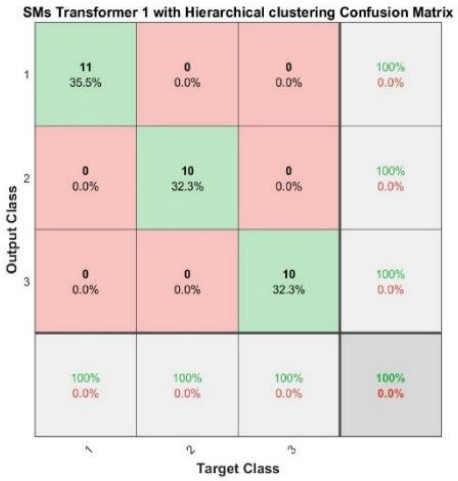

Figure 8.2 - Confusion plot for circuit 1 with full SM coverage (hierarchical clustering)

In Figure 8.3, the silhouette plot is obtained to assess the clustering quality of the groups obtained with hierarchical clustering, and it can be used to determine the optimal number of clusters for a given data set.

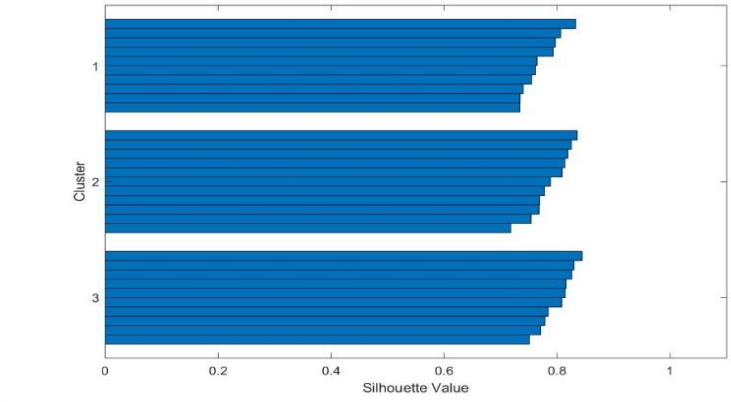

Figure 8.3 - Silhouette plot for circuit 1 (hierarchical clustering)

Since, in Analysis 1, these two algorithms are compared in order to find out the most optimal one. Thus, high silhouette values in the clusters are sought, and as it can be seen in Figure 8.3, the observations in the same cluster are highly correlated.

In addition, the graph shown Figure 8.4 represents how the half-hourly data for the time period of a month was grouped into three clusters for each phase:

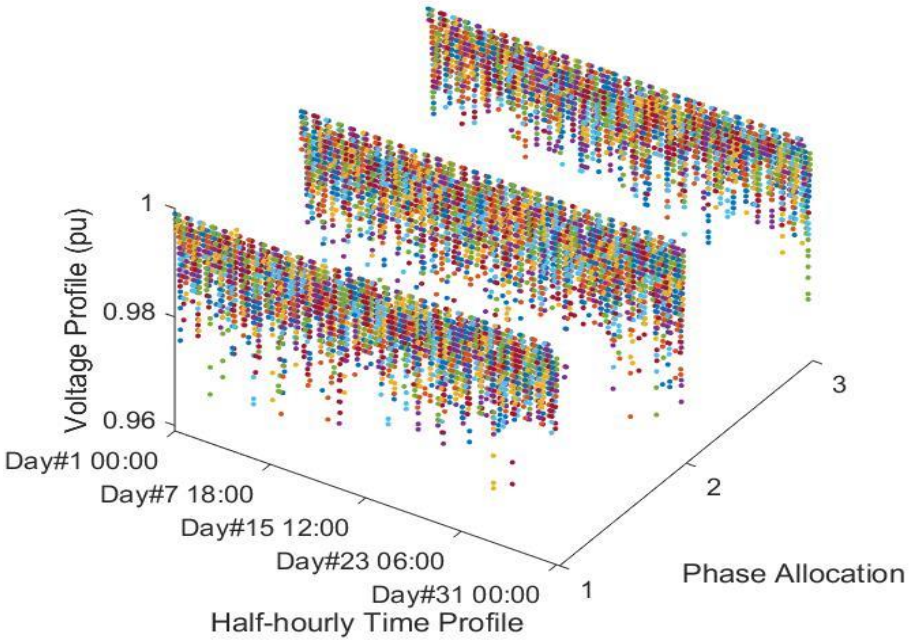

Figure 8.4 - SM voltage measurements grouped by phases (Hierarchical clustering)

8.1.2 K-means Clustering

In this section, the results of the different techniques are presented. These techniques are deployed with the objective of evaluating the accuracy and quality of the phase groups obtained with k-means clustering. The same techniques used in previous subsection are used, except for the dendrogram tree, which is a unique characteristic of hierarchical clustering.

Figure 8.5 portrays the confusion matrix that compares the phase allocation results obtained with k-means clustering with the target (real) results. A 100% level of success is also obtained for k-means clustering, in the same way as the results achieved by hierarchical clustering.

Figure 8.5 - Confusion matrix for circuit 1(k-means clustering)

Next, the results for the silhouette values are presented in Figure 8.6, which illustrates that the phase groups obtained with k-means clustering are largely correlated as well.

.

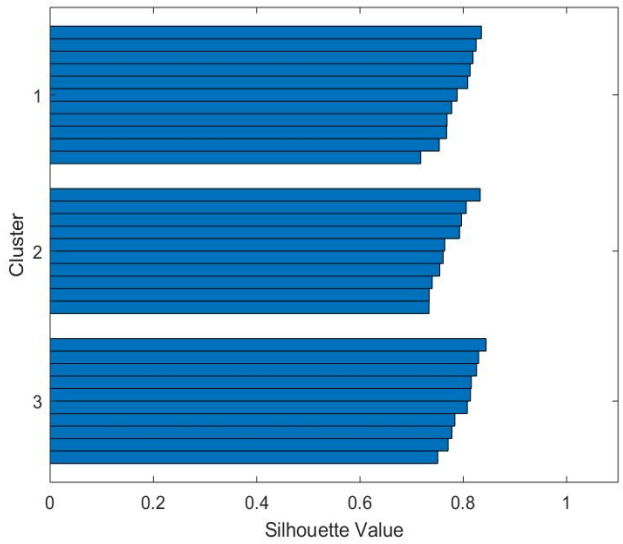

Figure 8.6 - Silhouette plot for circuit 1 (k-means clustering)

Lastly, in Figure 8.7, it is also represented how the voltage profiles were clustered in three-phase groups for data set corresponding to circuit 1 and the timespan of one month.

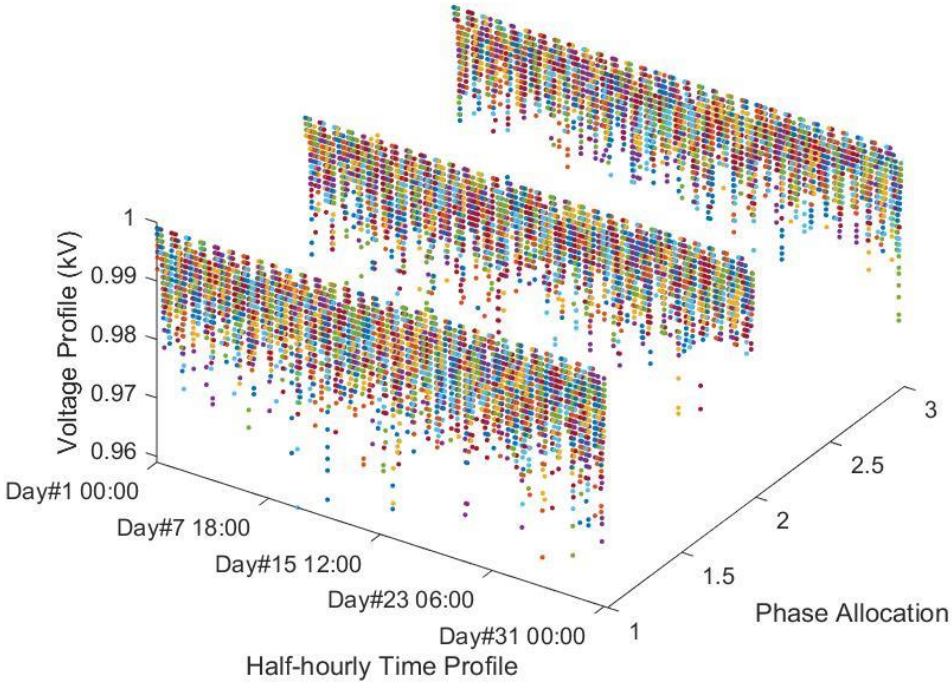

Figure 8.7 - SM voltage measurements grouped by phases (k-means clustering)

8.2 Reducing the Level of Penetration of SMs

As shown in the previous subsection, k-means and hierarchical clustering yielded equally accurate results for the phase allocation of the SMs in the LV network 1, for an ideal scenario where there is full coverage of SMs.

Therefore, the next step is to check if they would also be equally accurate for lower levels of penetration of SMs.

8.2.1 Method 1: SMs are Randomly Removal

Hereafter, the results obtained with methods 1 and 2, for decreasing levels of SMs penetration are shown:

Hierarchical clustering

Since multiple loops needed to be programmed in Matlab in order to obtain the results for the accuracy levels of k-means algorithm, different techniques from the ones used in section 8.1, are employed here.

It should be highlighted that since in this analysis SMs are removed on a random basis, the accuracy results may vary significantly. This is because the accuracy drops drastically when the algorithm is programmed to partition the data set into 3 groups, but all the SMs

connected to one (or two) of the phases have already been removed from the set. For this reason, the bar graphs shown in Figure 8.8, represent the accuracy levels obtained for levels of SM penetration from 0 to 100 per cent for different iterations.

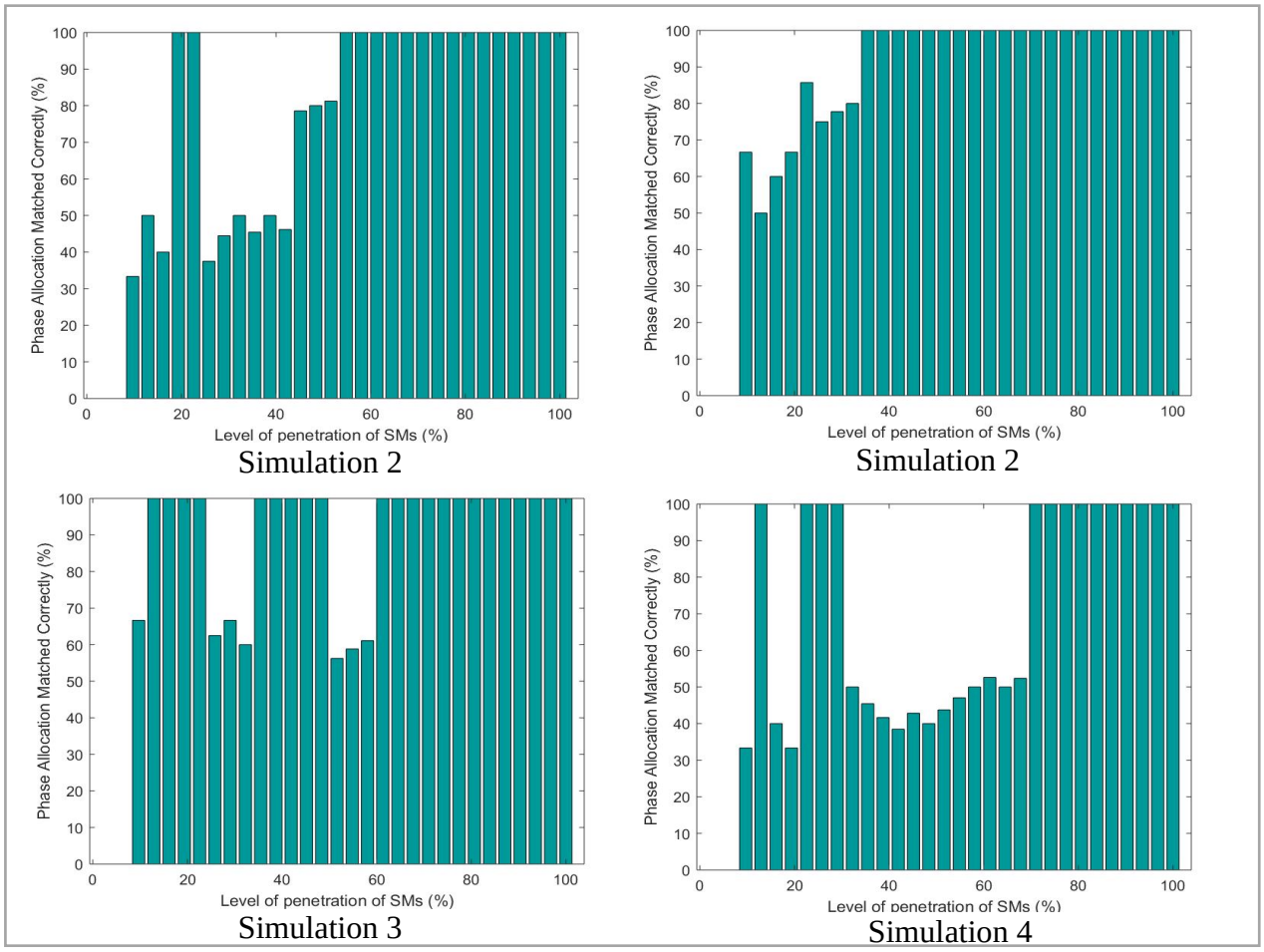

Figure 8.8 - Accuracy results for Method 1 (hierarchical clustering)

In addition to Figure 8.. in order to illustrate the level of presence of each phase as SMs are knocked out, that is, to show how the level of presence of SMs is being decreased, radar charts are presented for the four iterations.

These graphs allow to portray how multiple variables, which in this case are the level of presence of each variable and the total level of presence of SMs can affect the accuracy of the phase allocation algorithm.

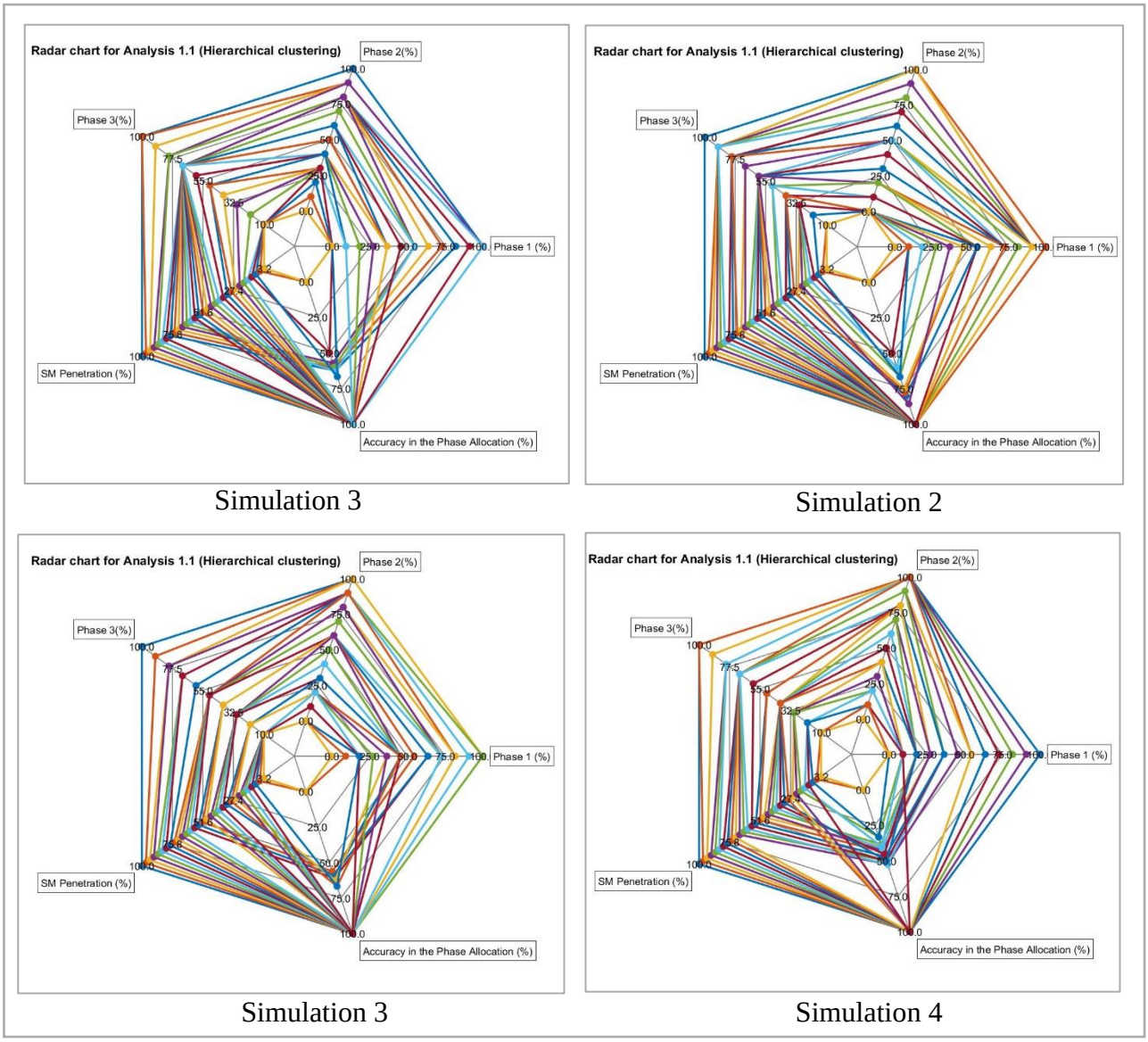

Figure 8.9 - Radar chart for Method 1 (k-means clustering)

It can be observed that both the level of penetration and the number of phases present in the network affect the effectiveness of the network.

K-means clustering

Similarly, in order to show the different outcomes that may result from randomly removing SMs, four simulations of the same k-means algorithm. Figure 8.. illustrates how these four simulations yielded somewhat varying results for different levels of SM penetration.

As it can be observed, and the accuracy results obtained are the highest up to levels of SM penetration of 10-20%, which shows this algorithm is very successful at correctly allocating phases for circuit 1 data set.

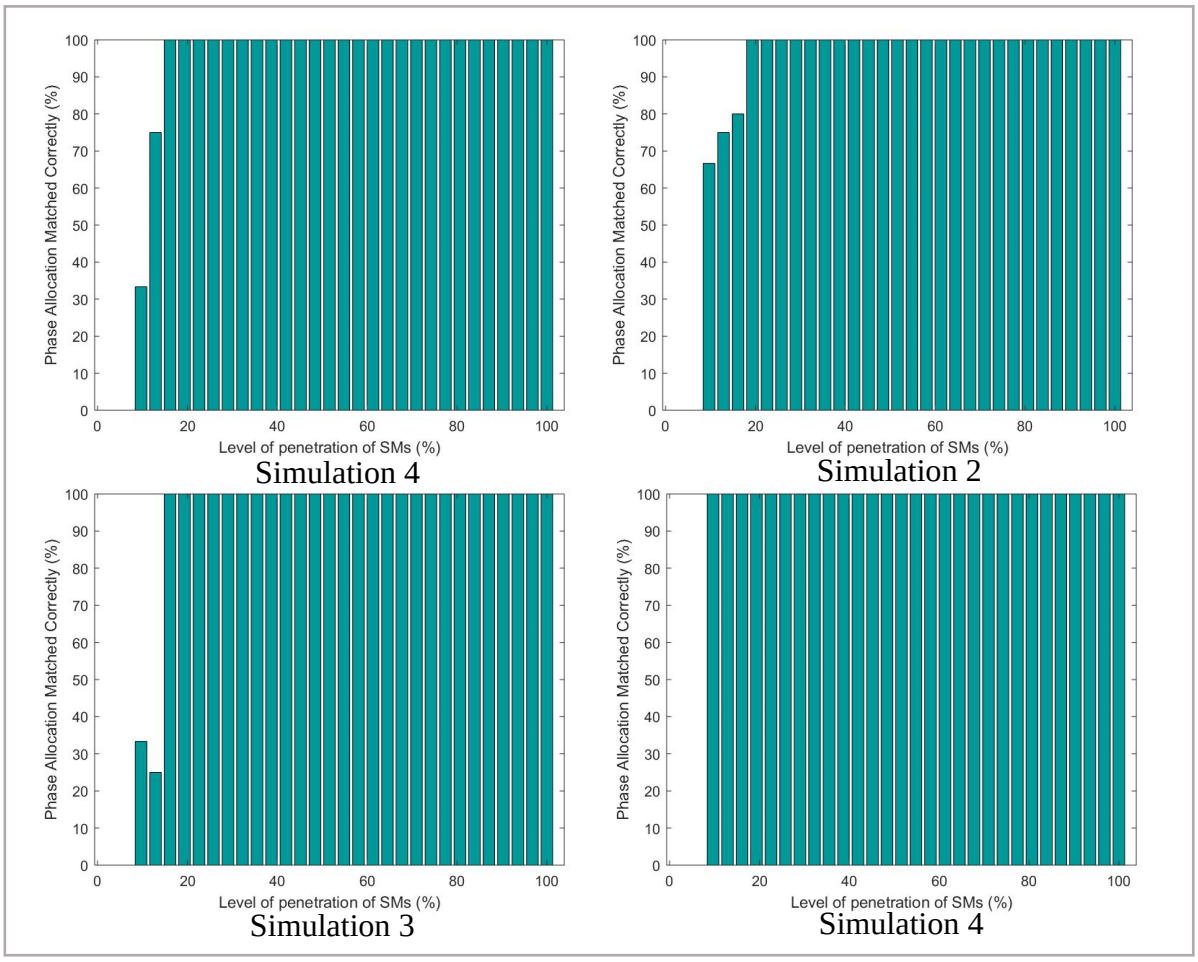

Figure 8.10 - Accuracy results for Method 1 (k-means clustering)

It is worth highlighting that more simulations were run, in order to check the variability of the results obtained with these methods, and the accuracy never was always 100 per cent for levels of SM penetration of 20 % or higher.

Also, in order to understand how different levels of phases presence and total percentage of SM in the network affected the results for the accuracy of k-means, the code for obtaining a radar plot was created in Matlab. The resulting graphs for the 4 simulations are presented in Figure 8.:

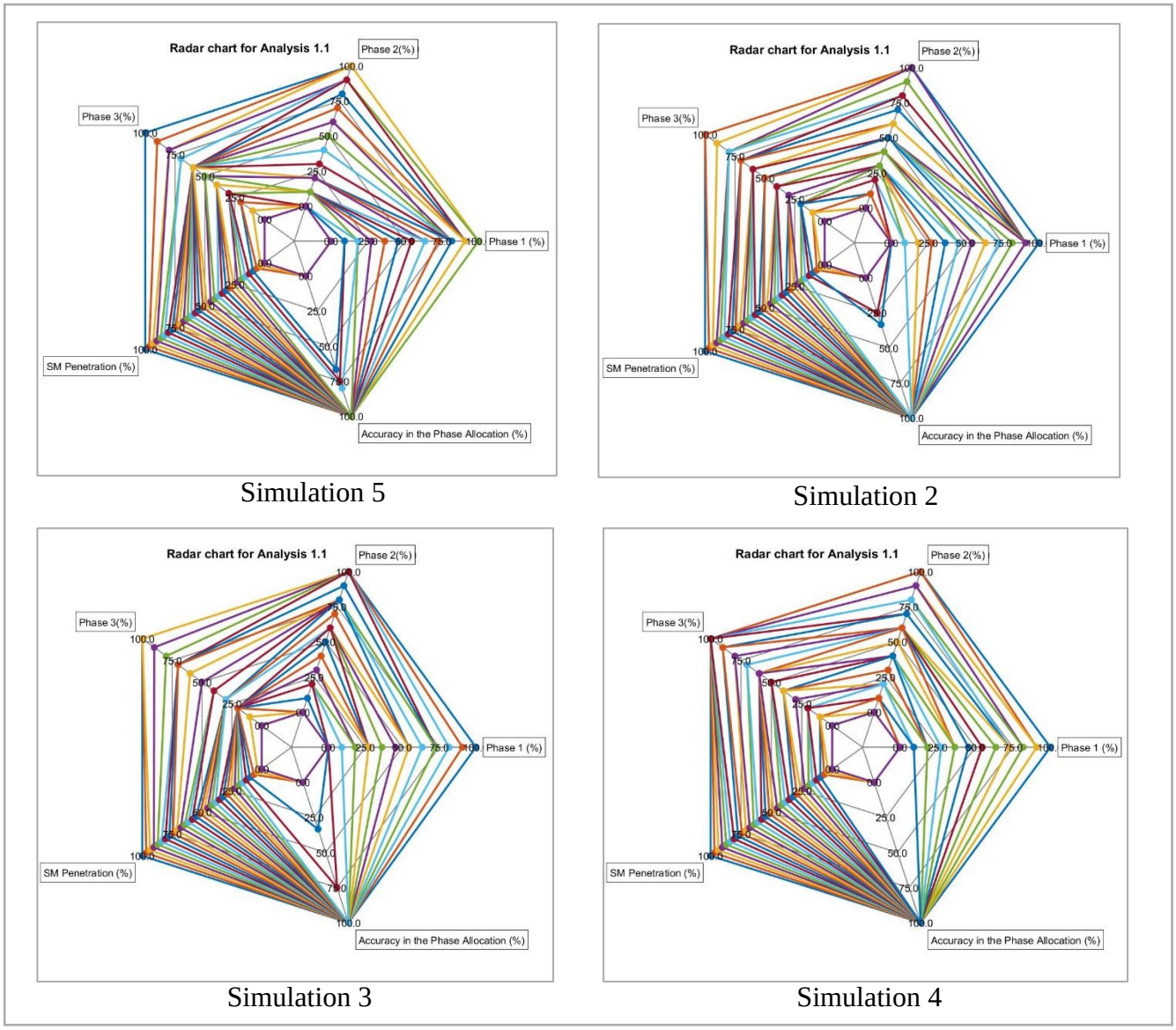

Figure 8.11 - Radar chart for Method 1 (k-means clustering)

The radar charts show that the accuracy of the k-means method is only decreased when there is no longer presence SMs connected to three different phases. This can be clearly observed in the iteration number 4 of Figure 8.: For this simulation, until the number of SMs analysed is 3 or less, the accuracy of the method is 100 per cent, since all the phases where present up to that moment.

8.2.2 Method 2: Orderly Removal of the SMs

In this method, SMs are removed in an orderly basis. In other words, SMs connected to phase 3 are knocked out first, then SMs connected to phase 2, and lastly, those connected to phase 1.

The objective of this method is to focus on how the accuracy of both methods is affected when only SMs connected two different phases are present in the data set. In addition to, the accuracy that can be achieved when there are SMs connected to only one phase.

Hierarchical clustering

Figure 8. shows the accuracy obtained for different levels of penetration with hierarchical clustering. Note that only SMs connected to phase 1 and 2 were still present, for percentages below 70%. And only phase 1 for percentages below 30%.

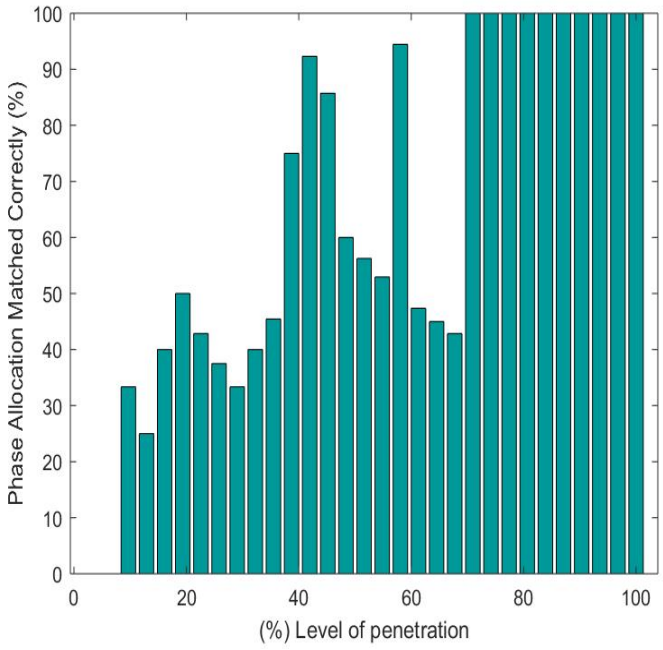

Figure 8.12 - Accuracy results for Method 2 (hierarchical clustering)

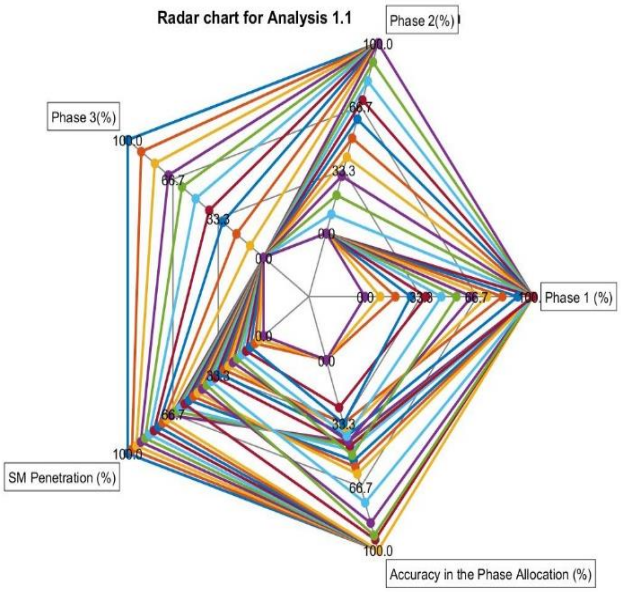

Figure 8.13 - Radar chart for Method 2 (hierarchical clustering)

This level of presence of different phases, as well as, the level of penetration of SM is depicted in the radar chart in Figure 8..

Both graphs show that only when there are three different phases present in the circuit, the algorithm is completely reliable for allocating phases,

K-means clustering

In the same way, the results for the orderly SM removal, are presented in both Figure 8.9 and Figure 8..

In this case, it can also be observed in both graphs, that only high levels of precision for the k-means clustering can be achieved when the number of phases corresponds to number of clusters for which the algorithm has been programmed to partition the observations.

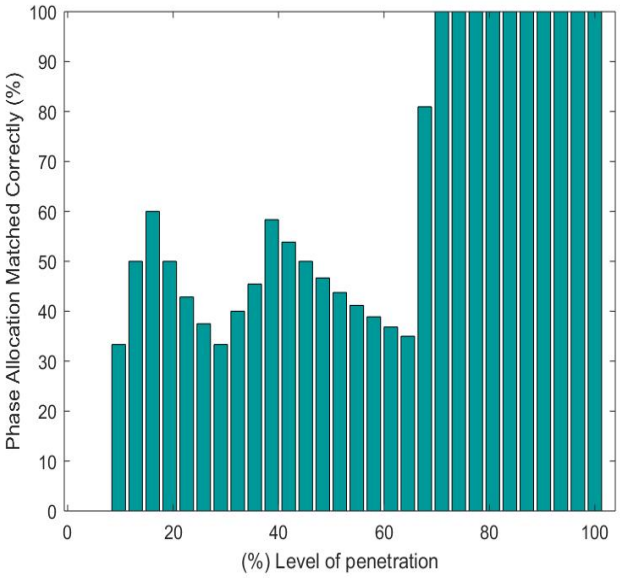

Figure 8.94 - Accuracy results for Method 2 (k-means clustering)

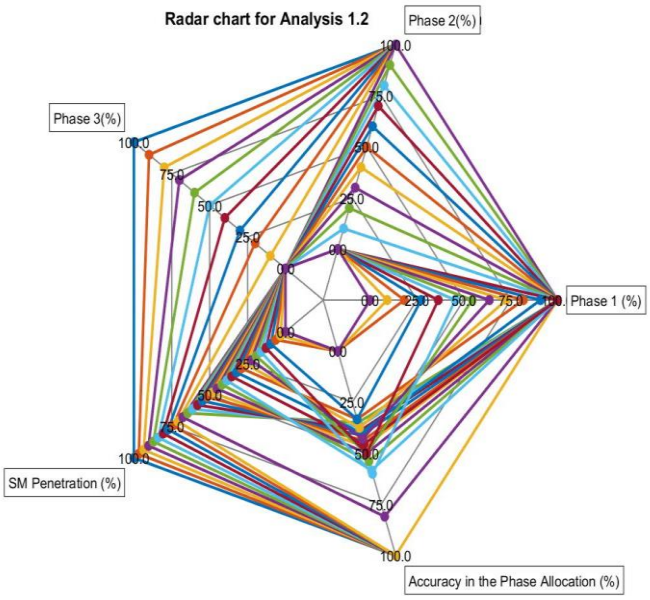

Figure 8.15 - Radar chart for Method 2 (k-means clustering)

8.3 Discussion on Analysis 1 Results

In this analysis, both algorithms were firstly tested for an ideal scenario in circuit 1, where there is full SM coverage. This was done largely for two main reasons:

- Firstly, to check that these methods could yield accurate phase identification results.
- And if so, to check the differences in the accuracy results obtained with both techniques.

After this first part of Analysis 1, it was proven that voltage correlation from SM connected to the same phase was enough to yield perfect phase matchings with both algorithms. What is more, no significant difference in the quality between the methods

can be observed in neither the confusion matrix nor the silhouette plot. As a consequence, these two methods were compared again, but this time simulating their accuracy in the phase identification for decreasing levels of penetrations.

The following table summarises the results obtained both for Method 1 (randomly SM removal) and Method 2 (SM connected to the same phase are removed orderly):

Table 1 - Summary of results for Analysis 1

	Average accuracy of the phase identification clusters	
	Hierarchical	K-means clustering
Method 1	~72%	~88 %
Method 2	~64%	~60%

K-means algorithm was proven to be a more accurate method for the clustering of phases based upon customer SM voltage measurements when Method 1 was employed. This means that k-means clustering was able to accurately partition the voltage measurements into clusters according to their corresponding phase connections even when very low levels of penetration were simulated.

On the contrary, k-means yielded way lower accuracy for the phase matching in Analysis 2. This means that this algorithm does not work as efficiently when the variety of different phase connections is reduced from three different phases two, as it can be concluded from the radar charts. And it decreases even more, when only SM connected to phase 1 are analysed. This was predictable since the major drawback of k-means clustering is that the number of clusters needs to be specified, and the k clusters value in these analyses was set to 3. Therefore, it is reasonable that its accuracy levels decreased when only two or one cluster(s) needed to be formed, but the algorithm partition the data set into three.

The effect of this phenomenon was not so apparent for k-means clustering with Method 1, since as the SM removal was performed on a random basis, the data set tended to have SM connected to all three phases for most of the loops.

However, hierarchical algorithm just proved to decrease its accuracy for identifying phases both when the level of total SM penetration is reduced, as well as when all the SMs connected to the same phase are knocked out. In fact, for both methods, when the level of SM penetration was lower than approximately 70% the predictability of the

accuracy that can be achieved with this method becomes somewhat chaotic. This because hierarchical clustering is only good at finding patterns in data when the data sets are large enough. Therefore, it can be considered that this is not a good method, if networks with low or intermediate levels of SM coverage should not be analysed with this approach. For this reason, even though hierarchical algorithm proved to be slightly better than k-means for Method 2, it was concluded that the drastic decrease in k-means accuracy be dealt with, by implementing an improvement in the k-means algorithm, which is presented in section 11.

Consequently, it was decided to proceed with the following analyses uniquely employing the k-means method.

9 Results for Analysis 2 and Discussion

In this section, the key outcomes from the analysis aimed at studying how the introduction of PV-DG in LV network can affect the phase identification problem, simulating decreasing levels of SM penetration with Method 1.

9.1 Results for Analysis 2

Among the given LV test networks, only circuits 4, 5 and 6 have available SM voltage data for both scenarios with a 50% level of PV penetration and with no PV penetration. For this reason, in this analysis, the k-means algorithm is tested in these circuits for both scenarios, with the goal of checking how the penetration of PVs can impact on the effectiveness of the method.

9.1.1 Scenario 1: No PV-DG

Firstly, the algorithm is applied to the voltage profiles generated for no penetration of PVs in the network.

These are the results for the following circuits:

LV circuit 4

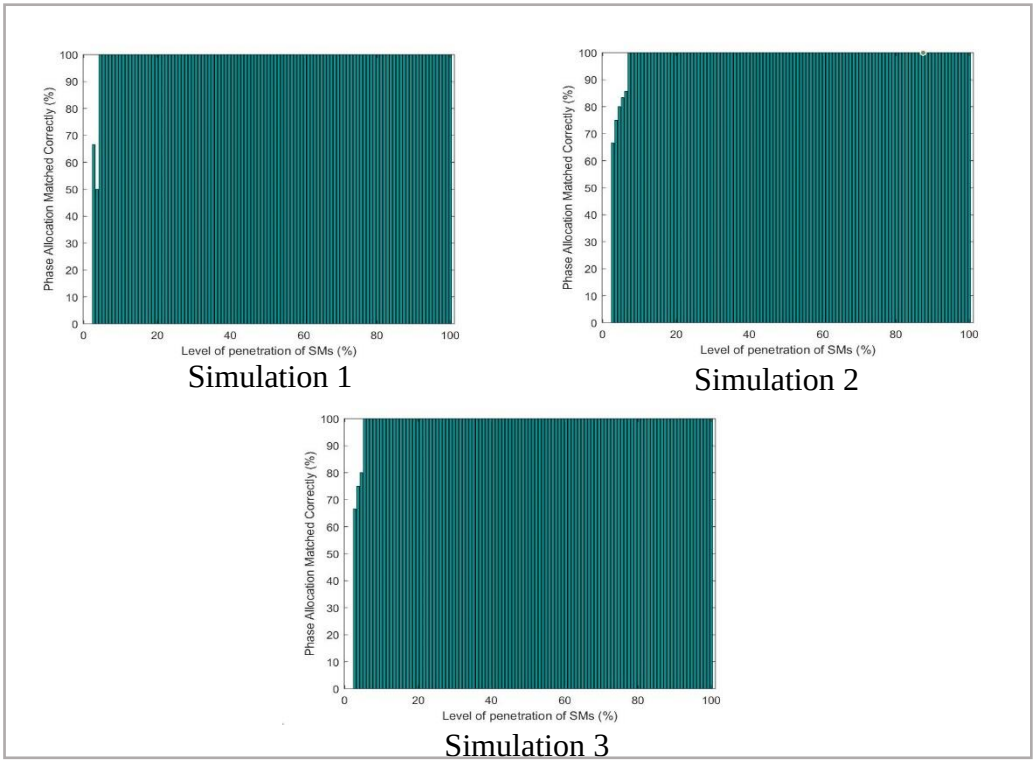

Figure 9.1 - Accuracy levels for method 1 and Analysis 2 (LV circuit 4 with no PV penetration)

These three graphs represent the accuracy levels for different levels of SM penetration.

Radar plots are used to illustrate the amount of SM connected to different phases in the network, as the total SM penetration decreases in each loop. Figure 9.4 shows the radar chart for these three simulations:

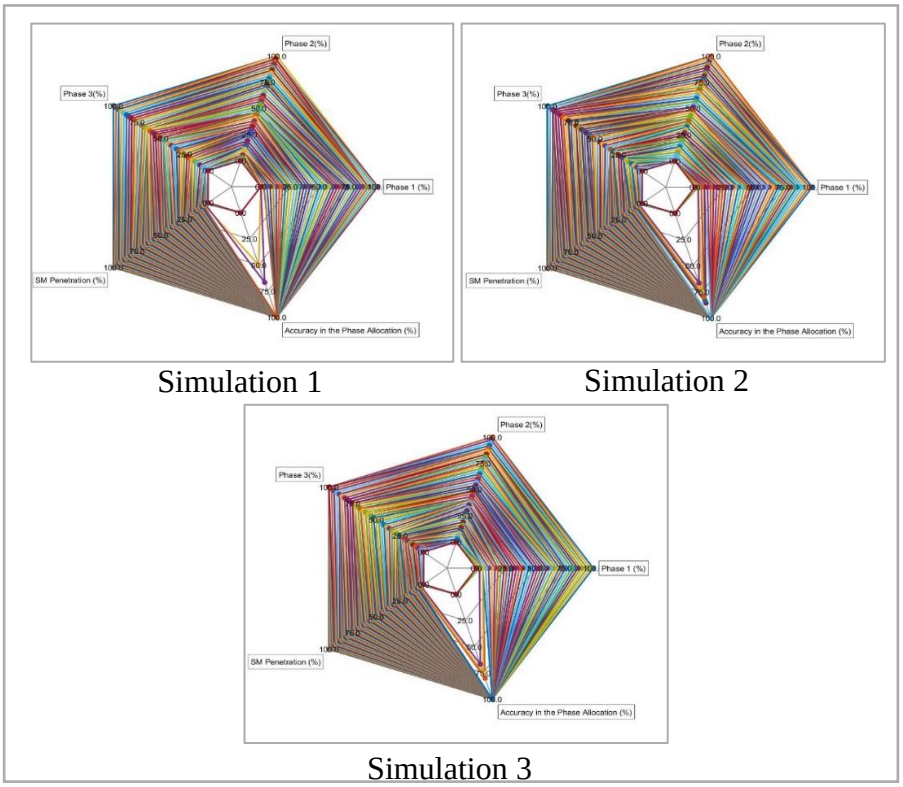

Figure 9.2 - Radar chart for method 1 and Analysis 2 (LV circuit 4 with no PV penetration)

LV circuit 5

The following graphs in Figure 9.3 depict the accuracy obtained for different levels of SM penetration for circuit 5.

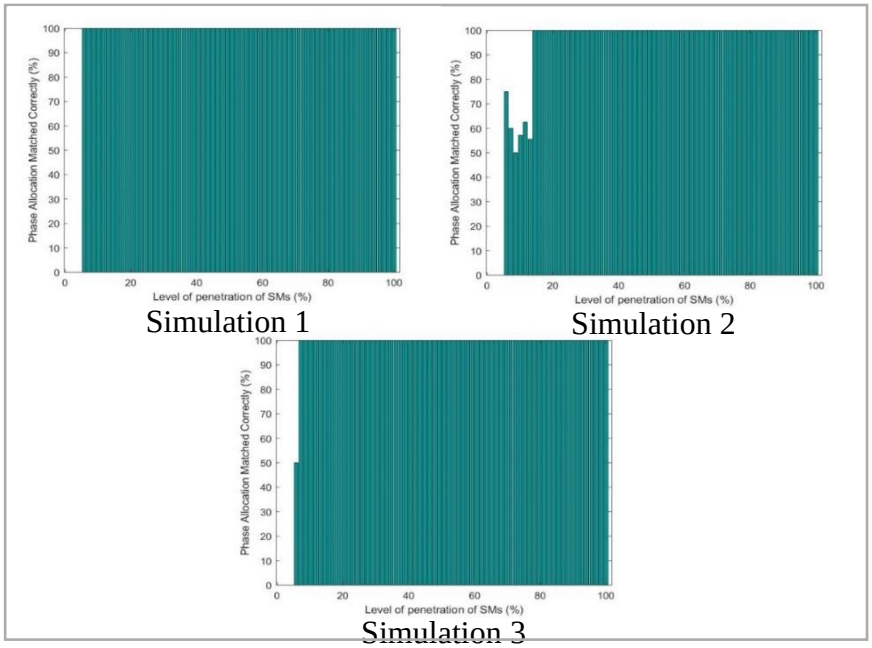

Figure 9.3 - Accuracy levels for method 1 and Analysis 2 (LV circuit 5 with no PV)

On the other hand, the radar charts represent the different percentage of phases and SM coverage, and how they affect the k-means accuracy in phase allocation.

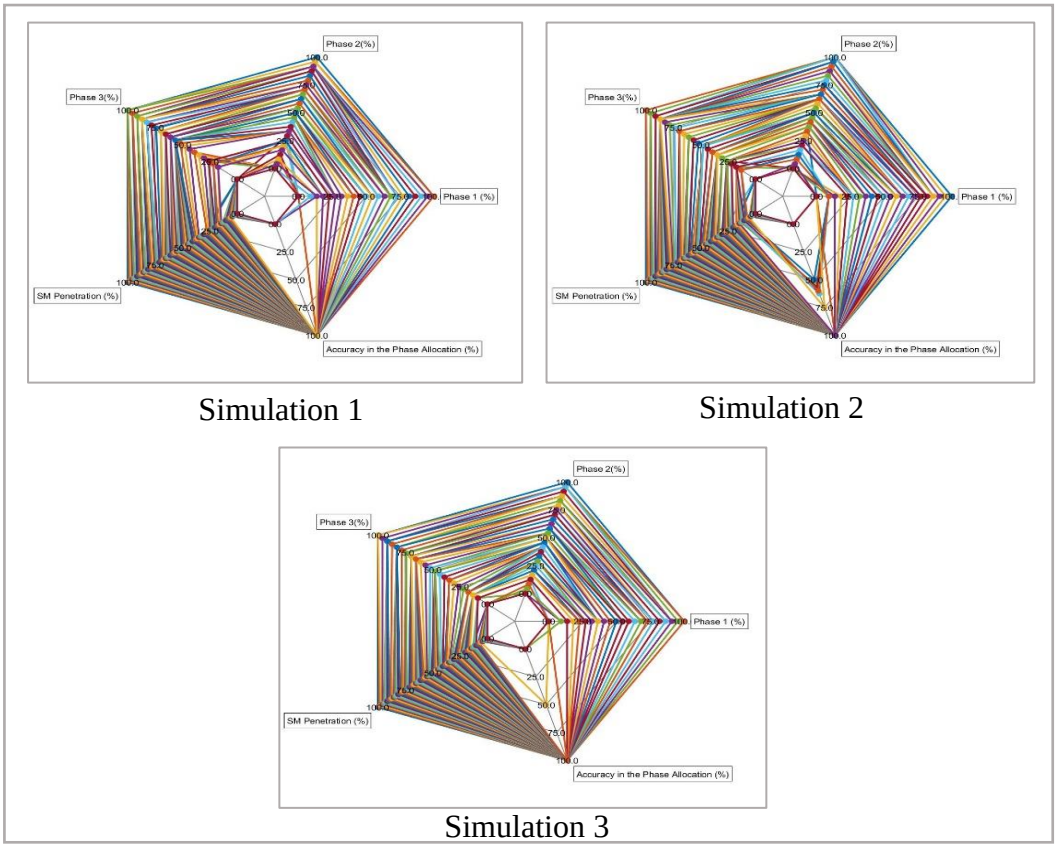

Figure 9.4 - Radar chart for method 1 and Analysis 2 (LV circuit 5 with no PV penetration)

LV circuit 6

Similarly, the results for the accuracy of the phase allocation method for SM penetration percentages ranging from 0 to 100, are presented in Figure 9.5:

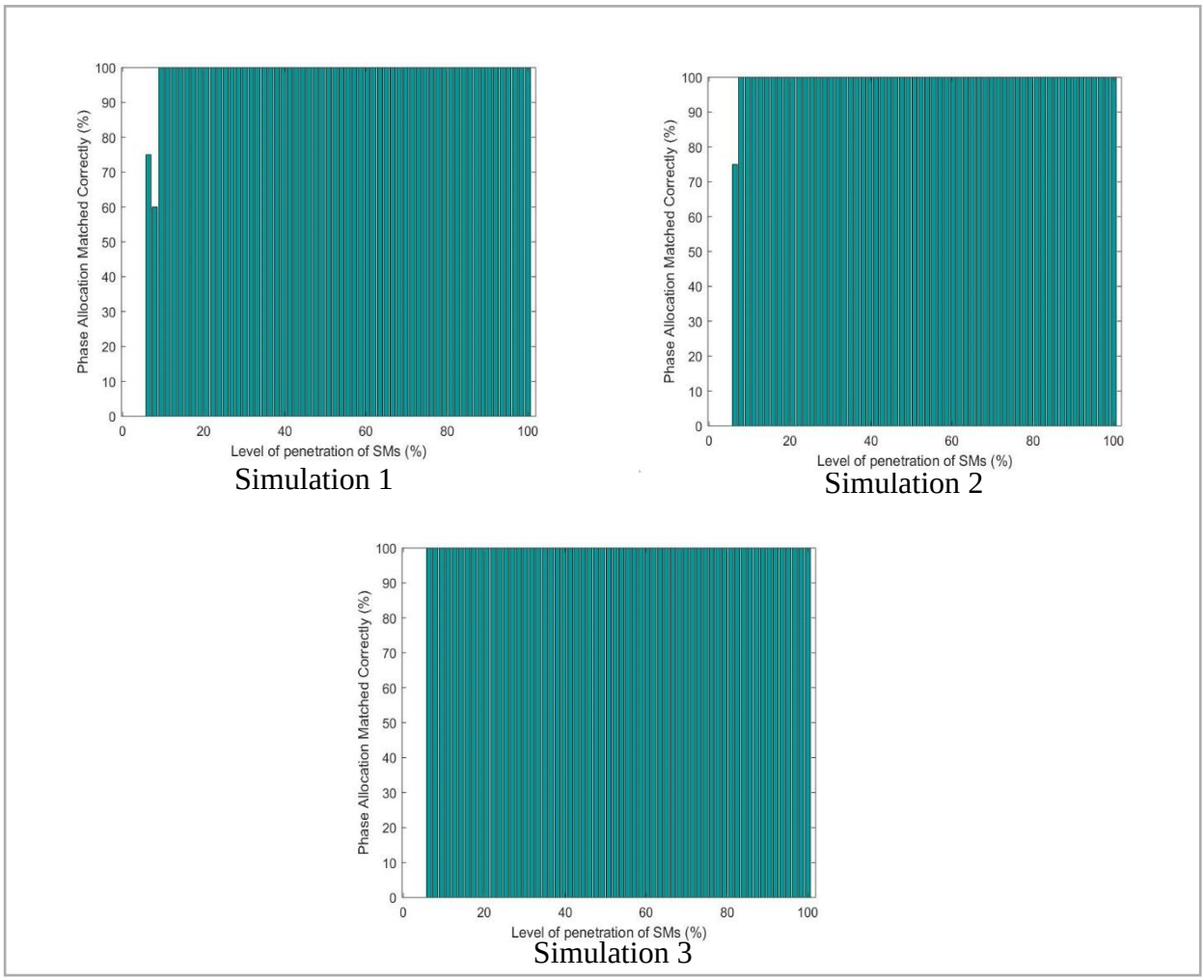

Figure 9.5 - Accuracy levels for method 1 and Analysis 2 (LV circuit 6 with no PV penetration)

And the next graphs, in Figure 9.6, allow to represent how the percentage of each phase and the SM penetration affected the accuracy of the phase allocation for the three simulations runned.

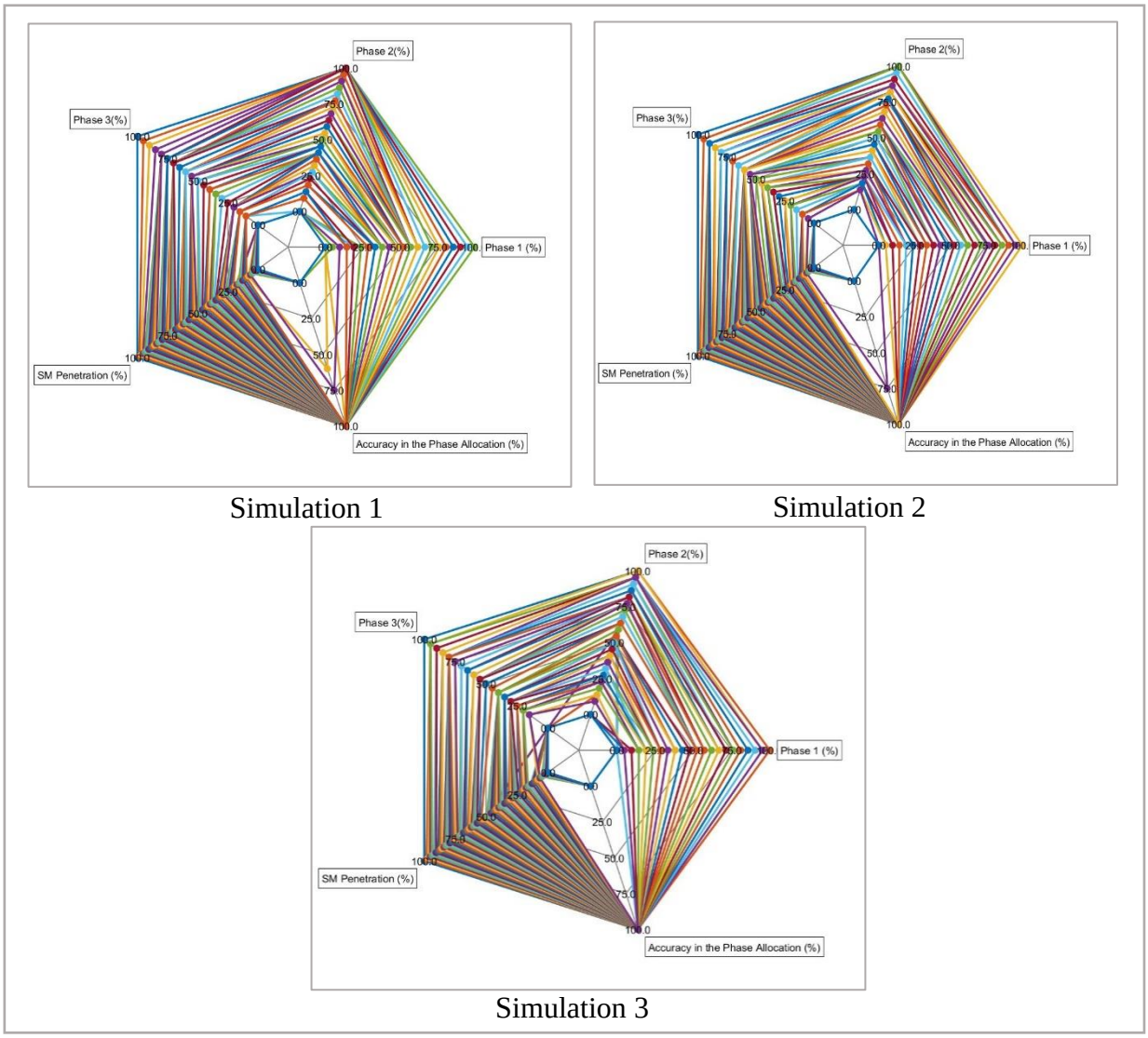

Figure 9.6 - Radar chart for Analysis 2 (LV circuit 6 with no PV penetration)

9.1.2 Scenario 2: 50% Level of PV-DG Penetration

Hereafter, the results for the same simulations as in section 9.2.1 are shown, with the particularity, that for this scenario the k-means algorithm is applied to a voltage data set generated from test network with a 50% level of PV penetration.

The results corresponding to circuits 4-5 are presented next:

LV circuit 4

The accuracy results from three different simulations are presented in Figure 9.7:

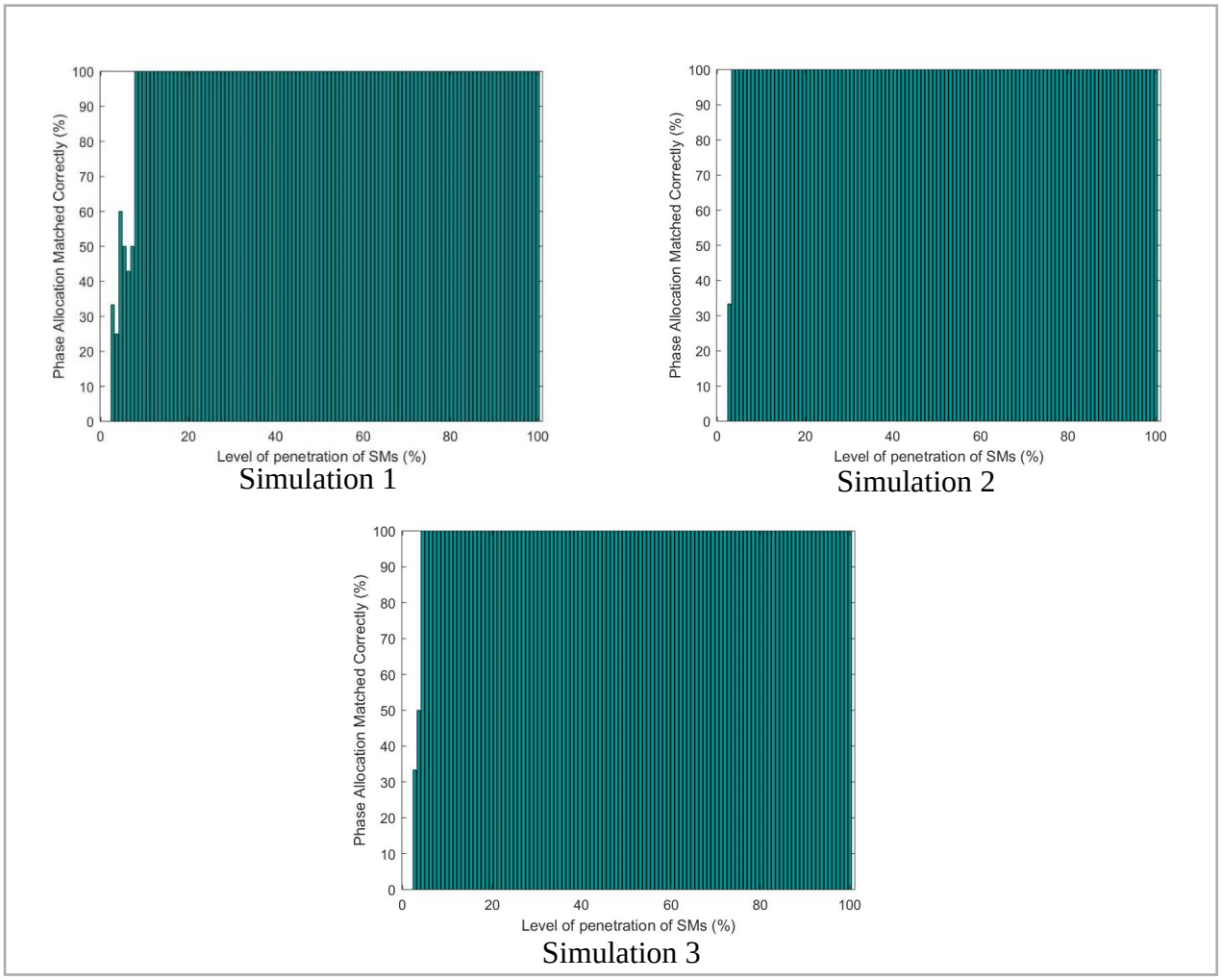

Figure 9.7 - Accuracy levels for method 1 and Analysis 2 (LV circuit 4 with 50% of PVs)

In addition to this, the radar chart resulting from comparing the several phase percentage and SM coverage with the accuracy obtained with k-means is shown in Figure 9.8

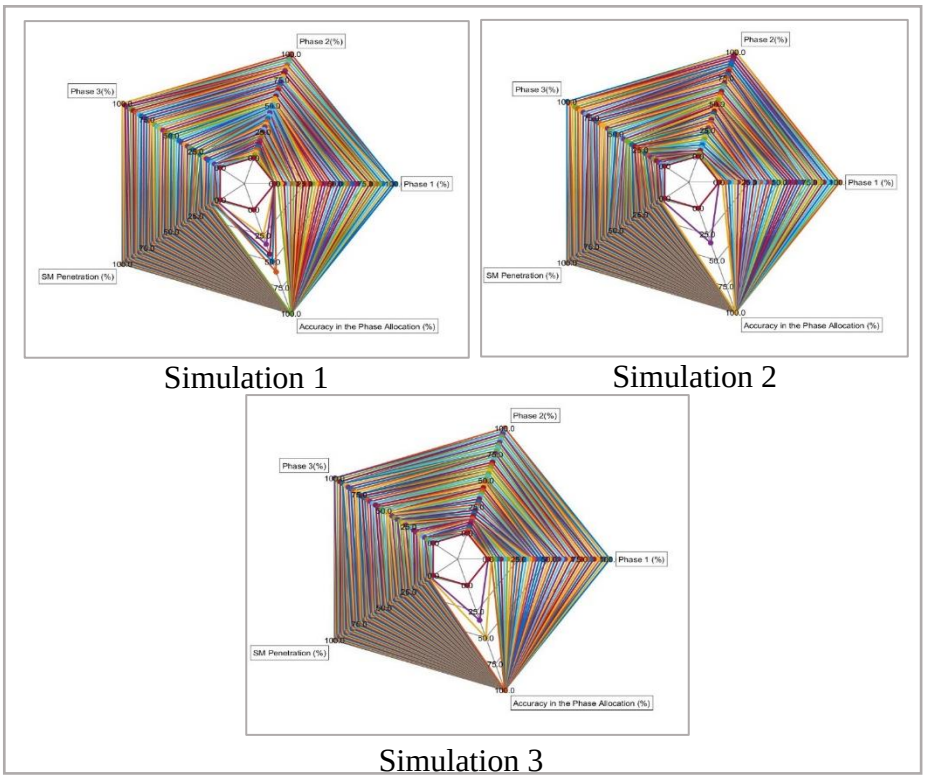

Figure 9.8 – Radar chart for method 1 and Analysis 2 (LV circuit 4 with 50% of PVs)

LV circuit 5

Figure 9.9 depicts the resulting accuracy degrees for PV penetration from 0 to 100 per cent:

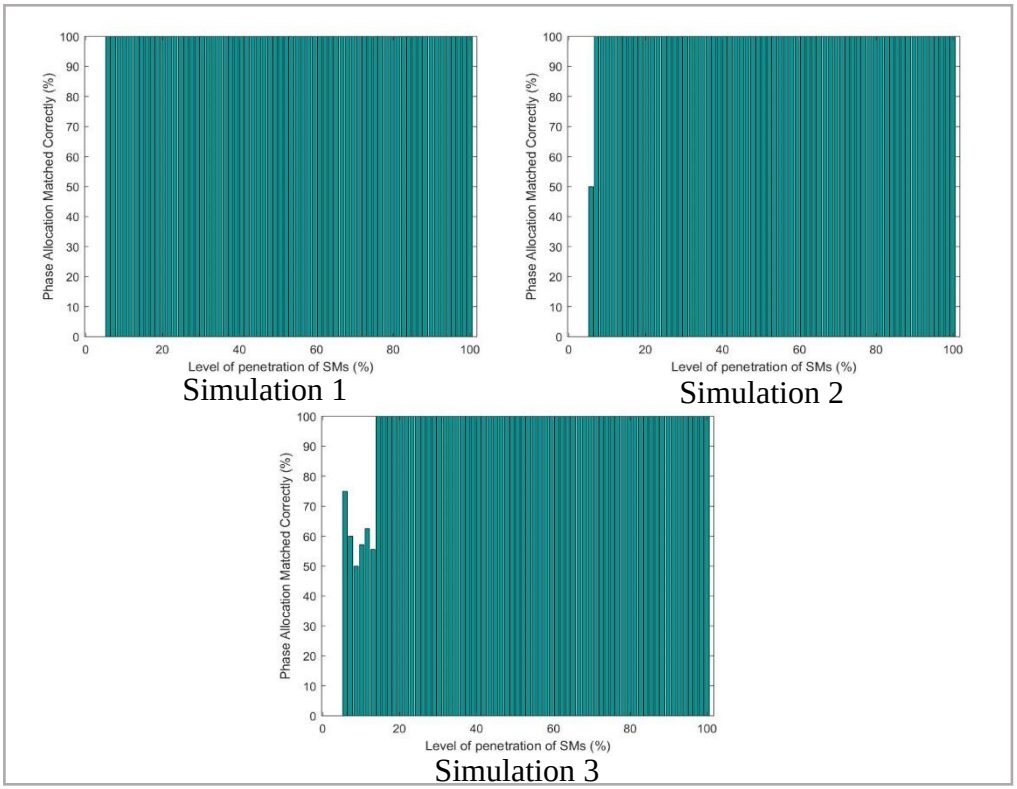

Figure 9.9 - Accuracy levels method 1 and Analysis 2 (LV circuit 5 with 50% of PVs)

Radar chart for the three simulations of the accuracy analysis are presented in Figure 9.10

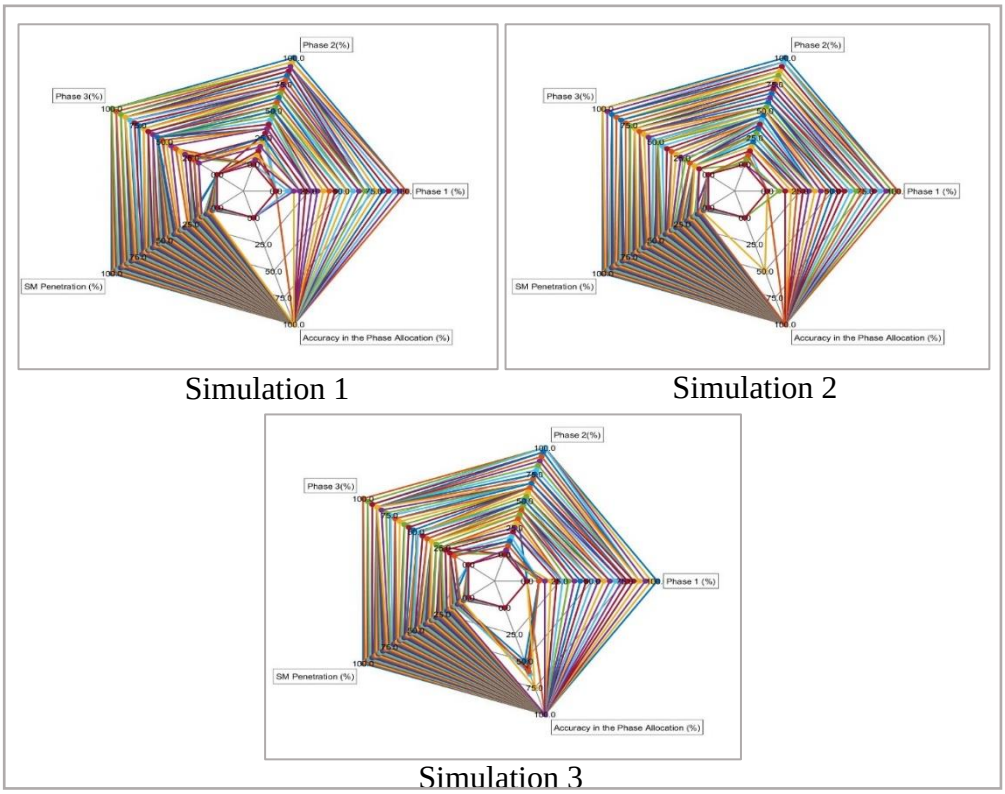

Figure 9.10 - Radar chart for method 1 and Analysis 2 (LV circuit 5 with 50% of PVs)

LV circuit 6

The accuracy results of the phase allocation for circuit and different levels of SM penetration are the following:

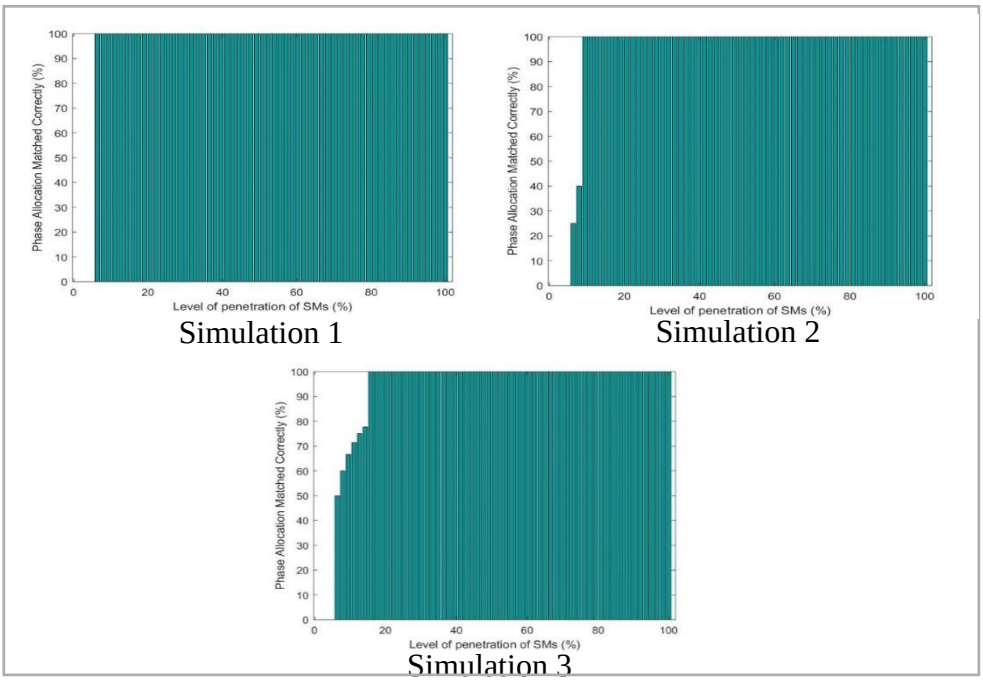

Figure 9.11 – Accuracy levels method 1 and Analysis 2 (LV circuit 6 with 50% of PVs)

Finally, the random removal of each phase percentage is represented in the radar chart, in Figure 9.11. In addition to how these varying amounts of phase and total number of SMs affect the accuracy of the results obtained with k-means.

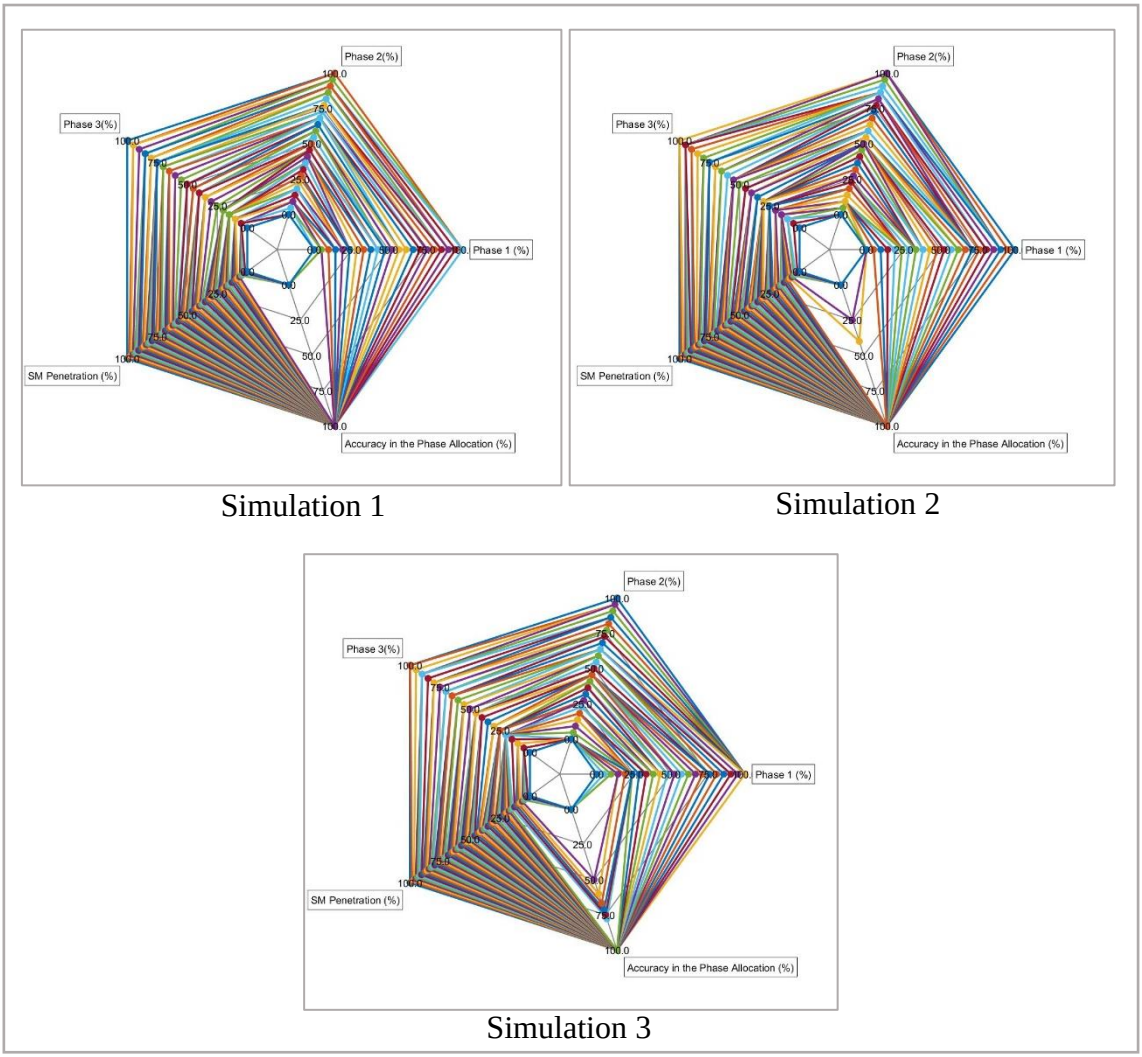

Figure 9.12 - Radar chart method 1 and Analysis 2 (LV circuit 6 with 50% of PVs)

9.2 Discussion on the Analysis 3 Results

The main findings of this analysis are the following:

- Likewise in Analysis 1, the results obtained for Analysis 2 enabled to prove that k-means clustering is also highly accurate for scenarios both with no PV penetration and with 50% of penetration, as long as, the number of different phase connection is known.

- The most significant contribution of this analysis is that it can be concluded that at least moderate PV penetration does not affect the accuracy of the phase

identification method. Since the results obtained for both scenarios were virtually the same.

- In fact, the decrease in the accuracy of the phase identification is given because of either really low SM penetration levels (~10-20%) or the complete removal of all the SM connected to one of the phases.

- In addition, the type of data used for this analysis is the average half-hourly data. Since the accuracy levels were quite similar to those obtained with the same method in Analysis 1, it can also be concluded that accuracy of k-means is not affected significantly by using either instantaneous or average voltage measurements.

10 Results for Analysis 3 and Discussion

This chapter provides the accuracy results obtained with the improved k-means approach, following Method 2, for simulating decreasing levels of SM penetration. Firstly, this section provides the comparison between the accuracy results that would be obtained for Circuits 1, 2 and 3 with the original k-means algorithm versus the improved one. This analysis is also performed for Circuits 4, 5 and 6, but for these circuits, it also compared the results obtained with and without PV penetration.

As explained in the method section, after analysing the results for method 2, it became apparent that k-means algorithm needed to be somewhat modified in order to overcome the fact it was programmed to partition the data set into 3 clusters, by introducing a gap evaluation before the clustering.

10.1 Results for the Three First LV Test Circuits

It should be noted that the data used for LV circuits 1, 2 and 3 is instantaneous half-hourly data.

10.1.1 LV Circuit 1

These are the accuracy levels of the phase matching obtained with the original k-means clustering vs the improved version with gap evaluation.

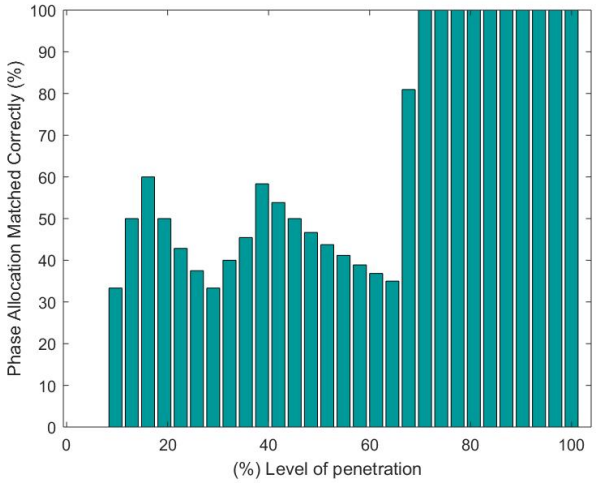

Figure 10.1 - Accuracy levels for circuit 1 (original k-means approach)

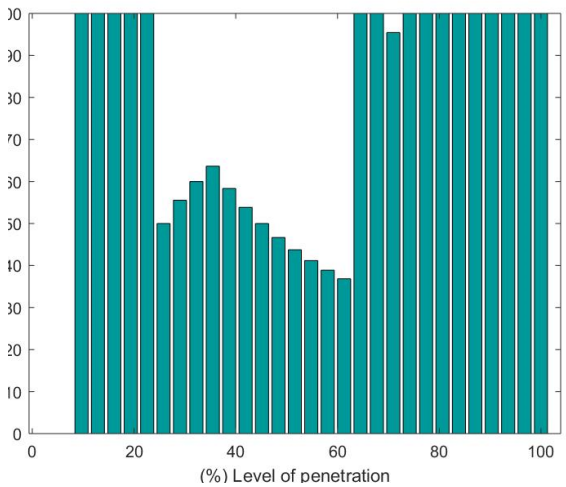

Figure 10.2- Accuracy levels for circuit 1(improved k-means approach)

10.1.2 LV Circuit 2

These are the results for circuit 2 with both k-means methods

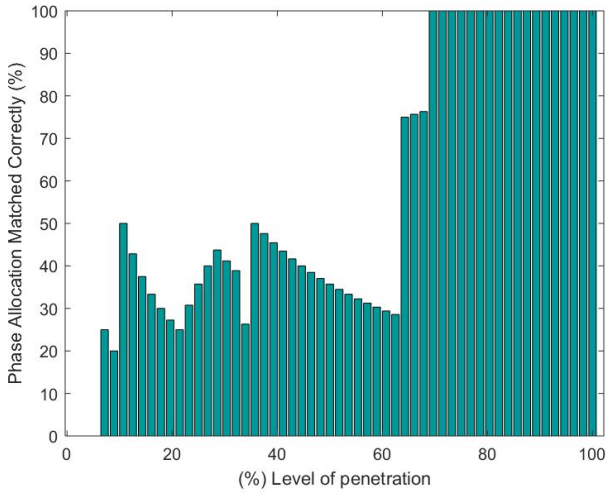

Figure 10.3 - Accuracy levels for circuit 2 (original k-means approach)

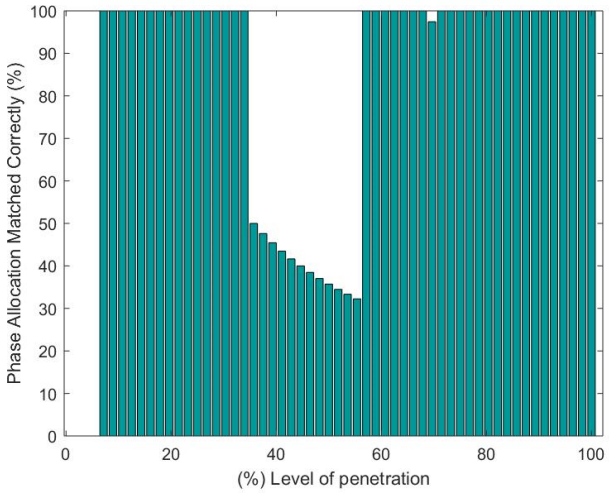

Figure 10.4 - Accuracy levels for circuit 1(improved k-means approach)

10.1.3 LV Circuit 3

The accuracy for the phase matching of both k-means approaches for circuit 3, can be observed in the following graphs:

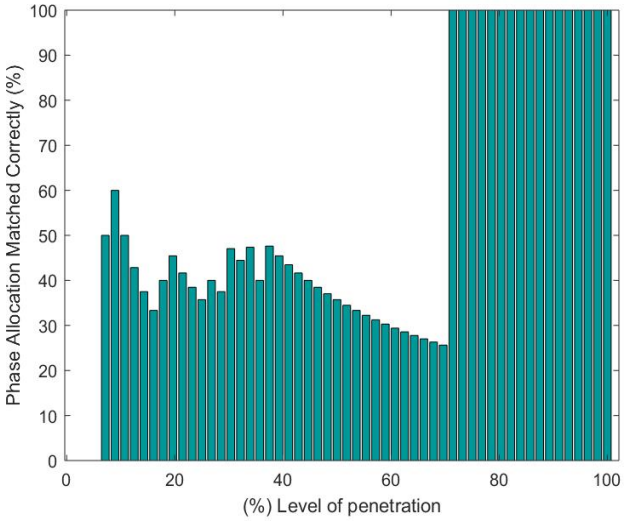

Figure 10.5 - Accuracy levels for circuit 3(original k-means approach)

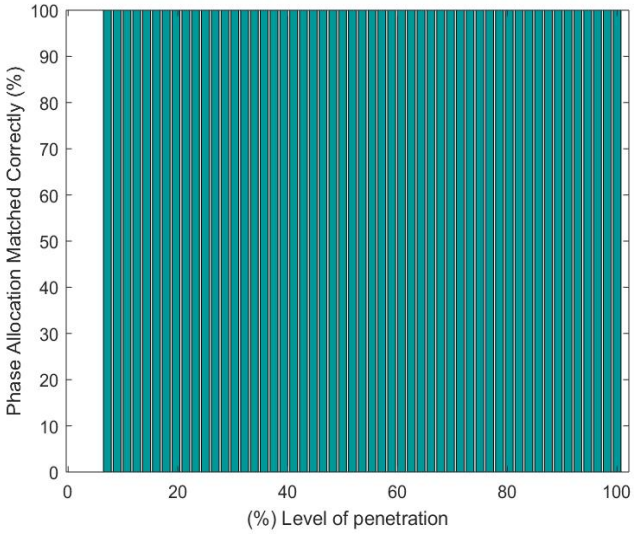

Figure 10.6 - Accuracy levels for circuit 3(improved k-means approach)

10.2 Results for the Last Three Circuits

The voltage profiles, in this case, are average half-hourly measurements. And note that the improvements in the accuracy levels are compared firstly, for the scenario with no PV-rollout, and then for a 50 per cent of PV penetration. The results for LV circuits 4, 5, and 6 are shown below:

10.2.1 Circuit 4

With no PV penetration

The following four graphs show the bar and radar charts showing how the accuracy level is affected by different levels of penetration (as well as, phase connections of SMs)

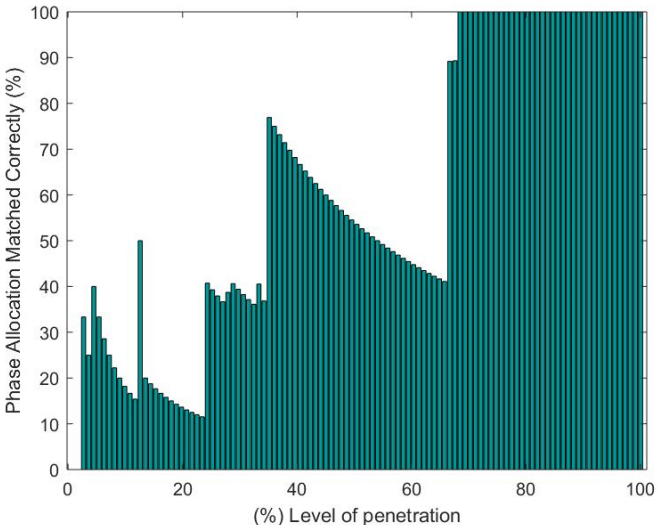

Figure 10.7 – Accuracy levels for no PV penetration in circuit 4 (original k-means approach)

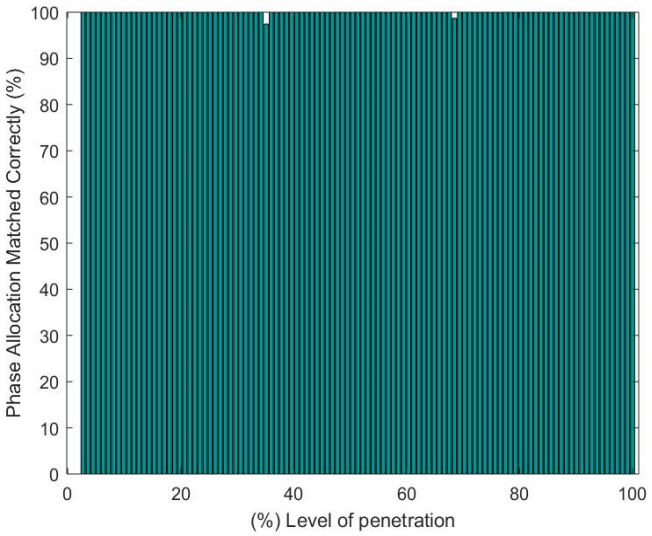

Figure 10.8 – Accuracy levels t for no PV penetration in circuit 4 with (improved k-means approach)

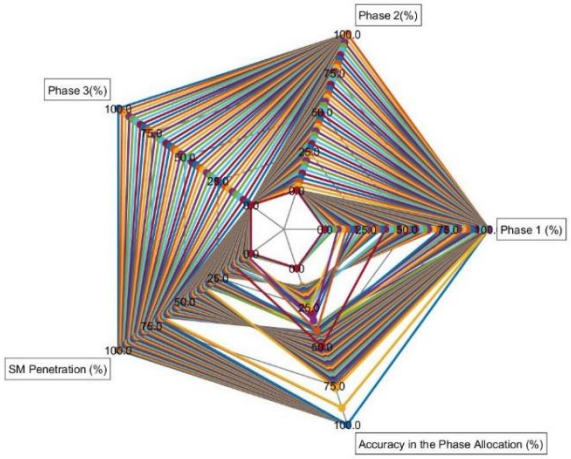

Figure 10.9 - Radar chart for circuit 4 with no PV-DG (original k-means approach)

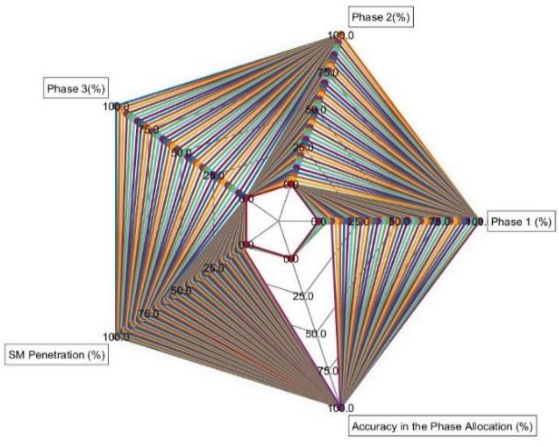

Figure 10.10- - Radar chart for circuit 4 with no PV-DG (improved k-means approach)

With 50% of PV penetration

The same results are presenteed in the following figures, but for the voltage profiles generated for a 50% of penetration of PVs in the circuit:

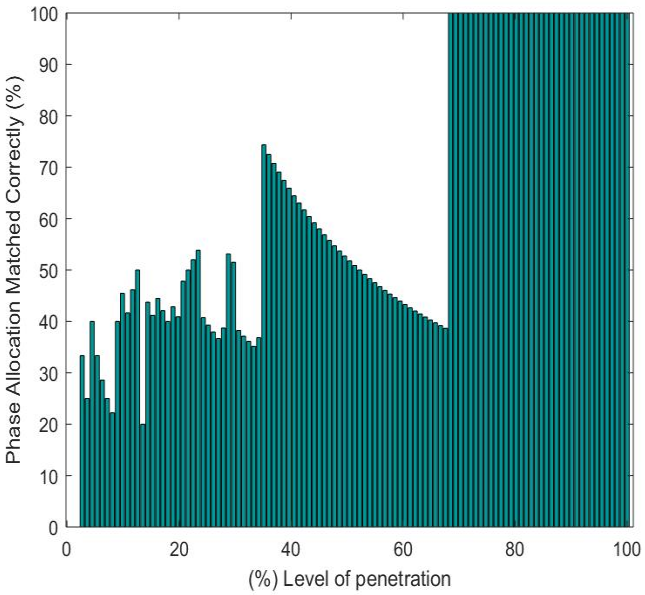

Figure 10.11 - Accuracy levels for a 50% of PV penetration in circuit 4(original k-means approach)

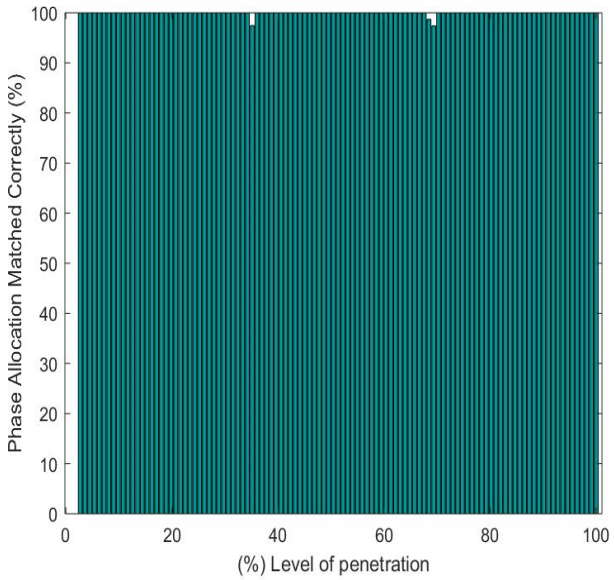

Figure 10.12 - Accuracy levels for a 50% of PV penetration in circuit 4 (improved k-means approach)

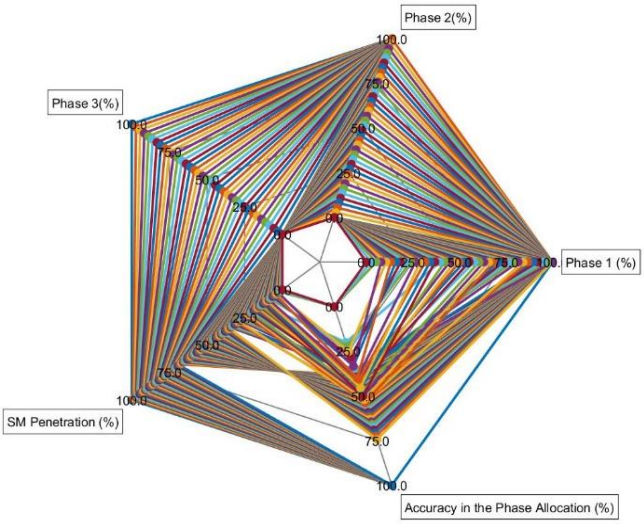

Figure 10.13 - Radar chart for circuit 4 with 50% of PV penetration (original k-means approach)

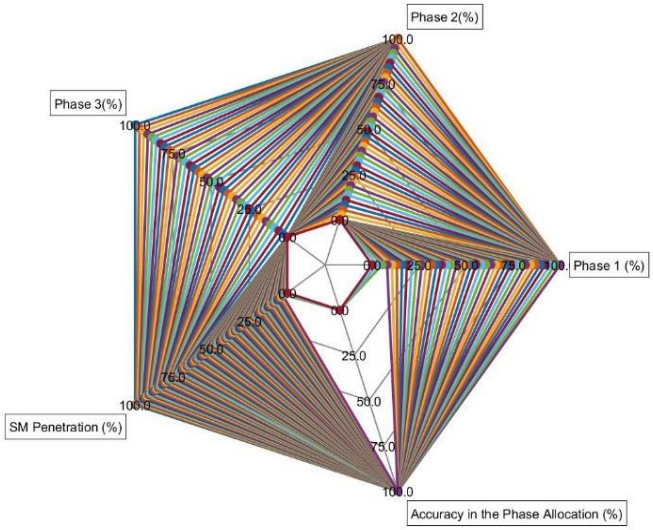

Figure 10.14 - Radar chart for circuit 4 with 50% of PV penetration (improved k-means approach)

10.2.2 Circuit 5

Hereafter, the phase identification results are presented for the half-hourly average voltage profiles from customer SM corresponding to circuit 5.

With no PV penetration

Firstly, the bar charts illustrating the accuracy levels for different SM penetration obtained with both algorithm are presented. Followed by their corresponding radar charts

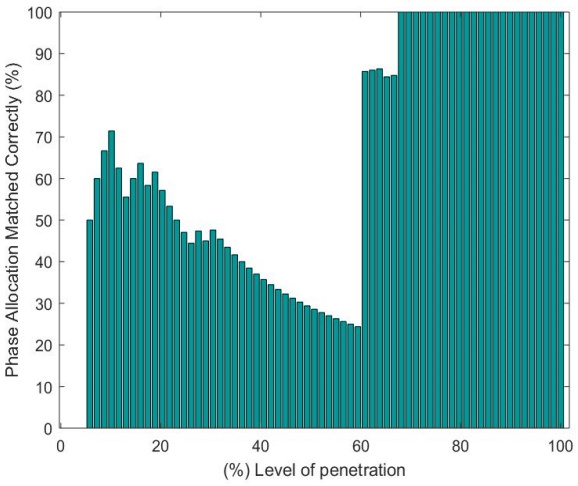

Figure 10.15 - Accuracy levels for no PV penetration in circuit 5 (original k-means approach)

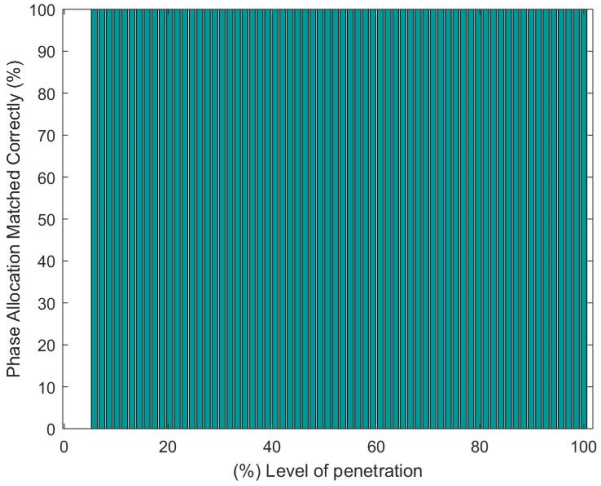

Figure 10.16 - Accuracy levels for no PV penetration in circuit 5 (improved k-means approach)

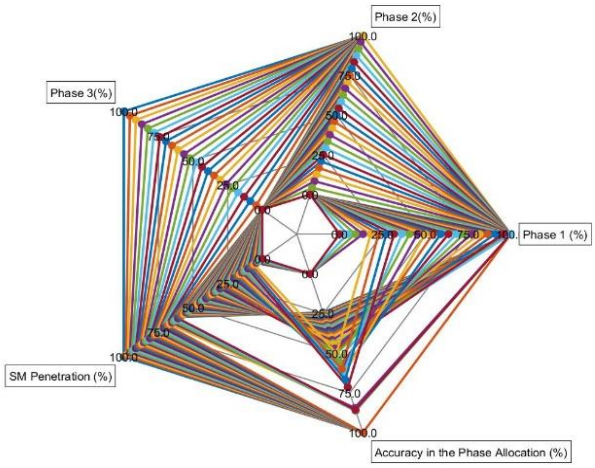

Figure 10.17 -Radar chart for circuit 5 (original k-means approach

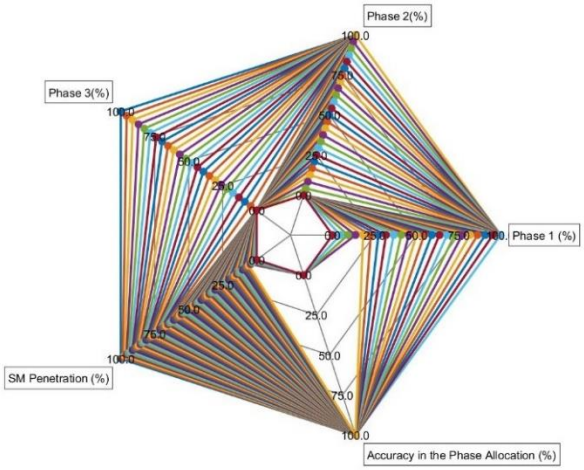

Figure 10.18 – Radar chart for circuit 5 with no PV-DG (improved (k-means approach)

With 50% of PV penetration

The following figures illustrate the accuracy results obtained with the original and the improved algorithm for circuit 5, and a moderate level of penetration of PVs.

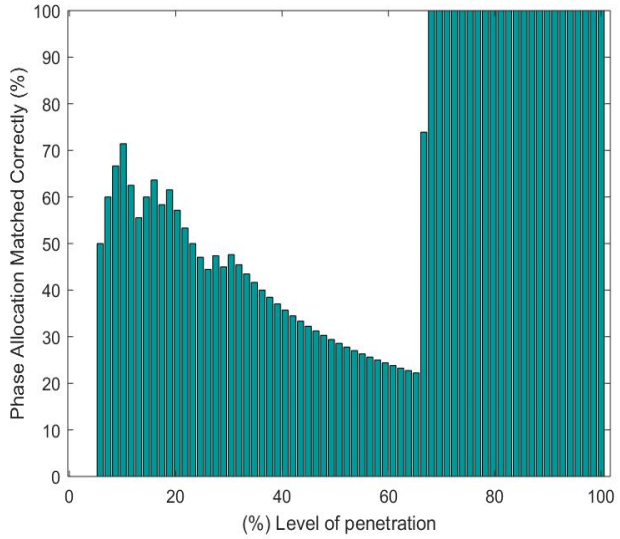

Figure 10.19 - Accuracy levels for 50% of PV penetration in circuit 5 (original k-means approach)

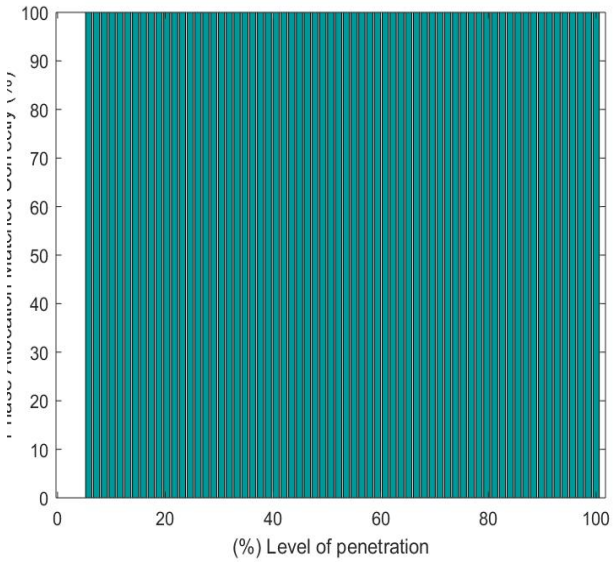

Figure 10.20 Accuracy levels for 50% of PV penetration in circuit 5 (improved k-means approach)

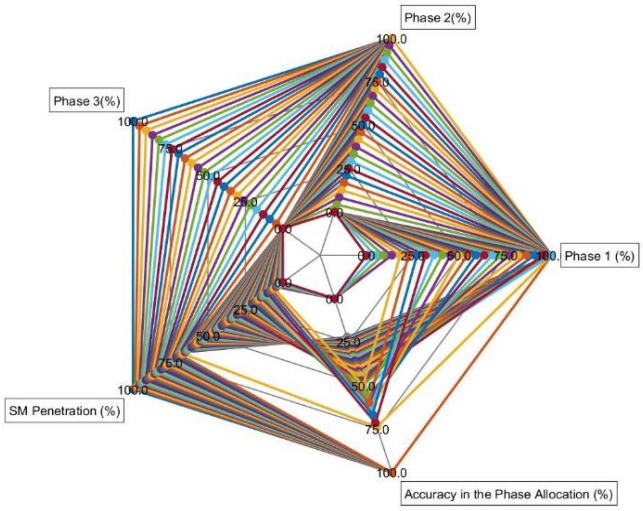

Figure 10.21 - Radar chart for circuit 5 with PV-DG (original k-means approach)

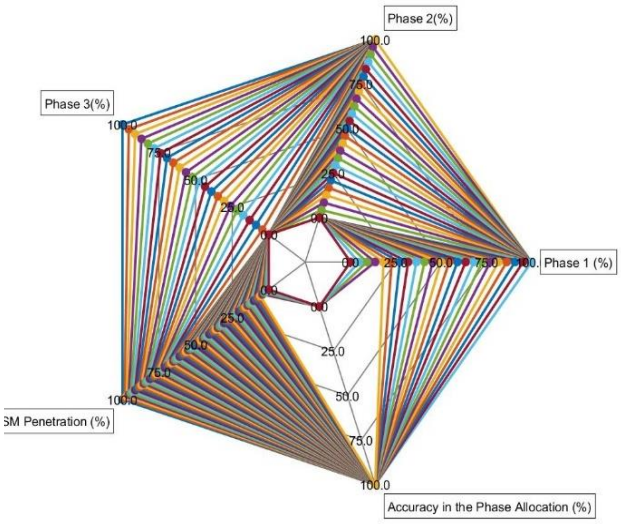

Figure 10.22 - Radar chart for circuit 5 with PV-DG (improved k-means approach)

10.2.3 Circuit 6

Finally, the results obtained with both algorithms for circuit 6 are shown next:

With no PV penetration

The scenario with no PV-DG in the LV network, presents the accuracy results for the different SM penetration levels, as shown in the following figures

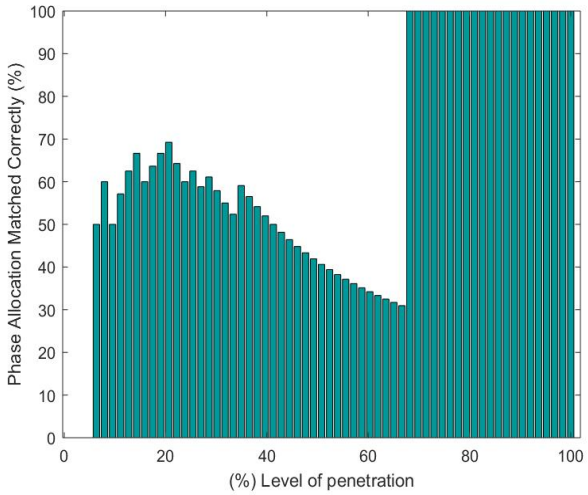

Figure 10.23 - Accuracy levels for no PV penetration in circuit 6 (original k-means approach)

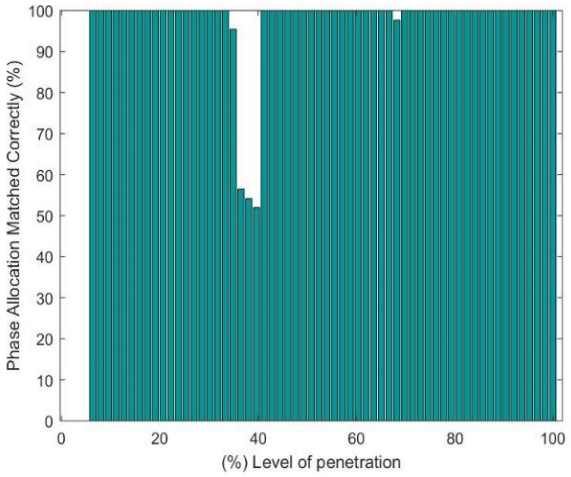

Figure 10.24 - Accuracy levels for no PV penetration in circuit 6 (improved k-means approach)

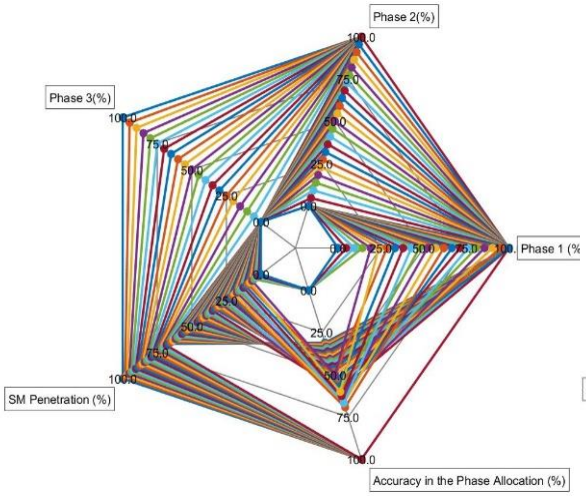

Figure 10.25 - - Radar chart for circuit 6 with no PV-DG (original k-means approach)

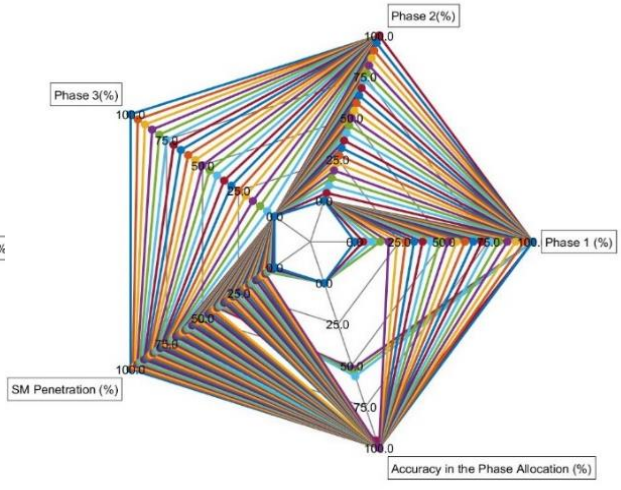

Figure 10.26 - Radar chart for circuit 6 with no PV-DG (improved k-means approach)

With 50% of PV penetration

Finally, these next four graphs represent the accuracy obtained with the original k-means clustering and the improved algorithm:

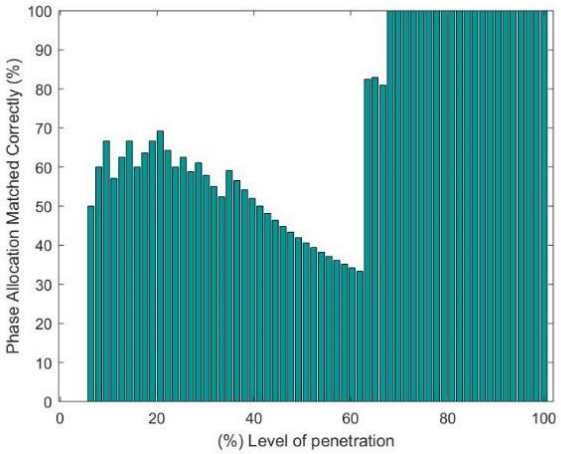

Figure 10.27 - Accuracy levels for 50% of PV penetration in circuit 6 (original k-means approach)

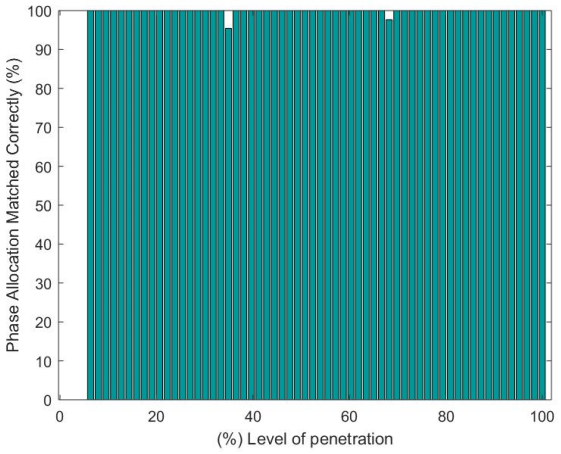

Figure 10.28 -Accuracy levels for 50% of PV penetration in circuit 6 (improved k-means approach)

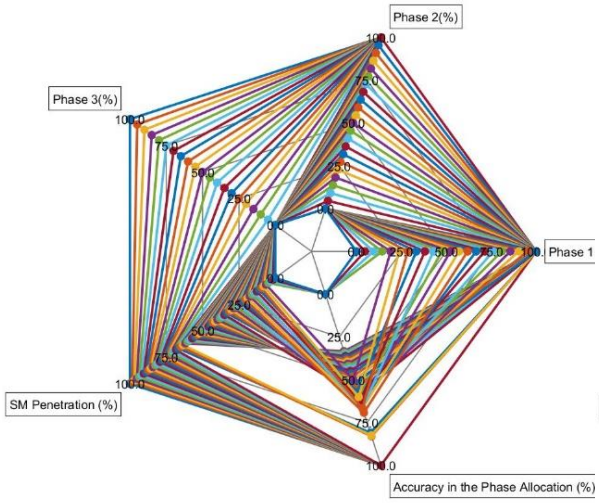

Figure 10.29 - Radar chart for circuit 6 with no PV-DG (original k-means approach)

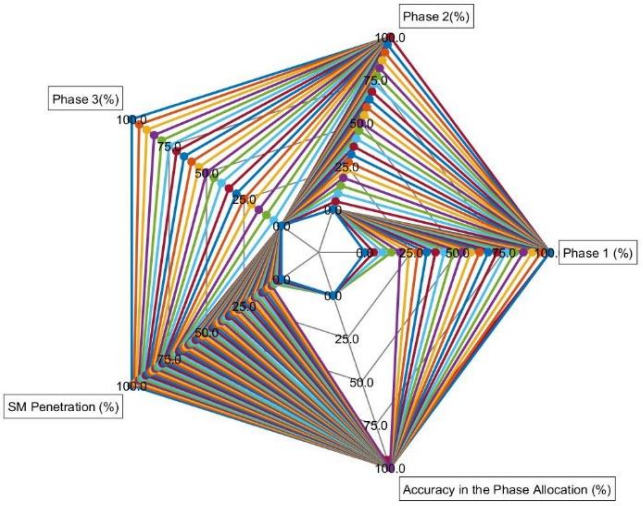

Figure 10.30 - Radar chart for circuit 6 with PV-DG (improved k-means approach)

10.3 Discussion on the Results from Analysis 3

As it was expected, higher accuracy levels for the correct matching of phases was obtained with the improved algorithm.

However, the accuracy obtained for circuits 1 and 2 is rather low, compared to the results obtained for Circuits 4, 5, 6. In particular, there is a drastic decrease for levels of penetration from 20-50%, this is for those levels, there were only SMs connected to phase 1 and 2, but the evaluation gap estimated that the optimal number of clusters was three. This means that the instantaneous half-hourly voltage profiles corresponding to customer SMs connected to phase 1 and 2 do not have high correlation factors different enough between clusters.

It is worth recalling that the clustering is performed in order that the SMs belonging to the same cluster have the highest correlations values among them, while the correlation values with the SMs belonging to other clusters is minimal. If clear clusters are not achieved, this is an indicator that correlation of instantaneous voltage measurements of customers connected to two different phases only is not enough to predict their phases correctly. These results may be surprising because the topology of circuits 1, 2, and 3, is simpler. However, it is also true that the amount of SM is more limited for these circuits, which reduces the accuracy of the gap evaluation criterion for this method. Wrong gap evaluation results lead to partition the data set into more (or less) clusters than the actual number; and therefore, triggering mistakes in the phase identification, and resulting in significantly lower results.

On the other hand, the evaluation gap criterion, and thus, the improved k-means algorithm proved to provide almost 100% average accuracy results for circuits 4, 5 and 6. This, therefore means, that the correlation between SMs connected to the same phase was high enough to form clearly defined clusters, even when the number of phase connection was reduced. Considering that the topology of these circuits is more complex, and that could have led to lower correlation values, it means that the success of the clustering may have resulted from evaluation half-hourly average data instead of instantaneous data. More research should be done in this area, but from analysing this six circuits, it can be concluded that average data present higher levels of correlation in the SMs connected to the same phase, and hence, this type of data is more desirable for achieving optimal results with the proposed methods.

In addition, it is remarkable that such high levels of accuracy could be obtained for circuits with a considerable level of branching and number of consumers, and most importantly, with a 50 per cent of PV penetration.

Consequently, the main contribution of this thesis, therefore, is that accurate phase identification can be achieved even in data-constrained scenarios with moderate PV-DG scenarios with a variation of the k-means clustering technique.

11 Conclusions

This thesis investigated the feasibility of implementing a phase identification approach for the given highly constrained scenarios. After an extensive literature review, two methods that had already been used successfully for solving the phase identification problem were tested in the LV networks under study.

From this analysis, it was concluded that hierarchical clustering is less effective than the k-means clustering technique. The accuracy of the phase allocation results with the hierarchical technique was more seriously affected by decreasing levels of penetration. This was expected from the literature review since hierarchical clustering is thought to work less efficiently for large data sets.

Therefore, the remaining analyses were performed uniquely with the k-means clustering, which is a method that has been broadly used in literature. However, it had never been used for analysing voltage measurements from the customers' SMs uniquely. And although it was also learnt from literature that voltage profiles from SMs connected to the same phase are correlated, it was unknown up to what point that correlation would allow to correctly partition the measurement into clusters, identifying the phases of the SMs. In order to test this, two more analyses were performed, with the objective of determining how different scenarios and factors could affect the accuracy of the phases allocated with the k-means algorithm.

From the whole three analysis, it can be concluded that the following are the main contributions of this thesis:

- Hierarchical clustering requires levels of SM penetration around 70% or higher in LV circuits similar to the ones studied in this thesis. That is, with limited data sets. Further research is required in order to analyse how this limitation could be tackled.

- The original k-means clustering demonstrated to be rather limited when the number of different phases to which customer's SMs are connected to are unknown. However, accuracies close to 100% were achieved even for low levels of penetration, when the number of clusters to be formed were known (i.e. the number of different phase connections)

- An improvement for this original k-means clustering technique was introduced with highly accurate phase identification results when average voltage profiles were analysed. Therefore, concluding that average voltage measurements are preferable over instantaneous ones, when k-means clustering techniques are to be applied.

- The presence of a moderate percentage of PVs on the rooftops of the LV networks did not prove to disturb in any way the results obtained with the k-means method. Meaning that, at least moderate penetration levels of PVs, do not alter the correlation factors present in voltage measurements of customers connected to the same phase. Once again, further research is needed in order to investigate how higher levels of PV penetration can affect these correlation factors, and therefore, the quality of the phase identification by clustering techniques.

Overall, it can be concluded that the objectives presented in section 1.2, have been fulfilled, but this topic resulted to be a very interesting and important topic, which would need further research in order to verify that the results achieved with the proposed methods are accurate enough for other scenarios and LV networks. Future lines of research are discussed in the next chapter.

12 Future Lines of Work

The following section details suggestions for future work to be carried out to address the limitations of this thesis.

Although satisfactory results were achieved in this thesis considering how challenging the scenarios analysed were, this topic could have been further developed in many ways. Phase identification-related investigations are an exciting topic, and it will become more significant as the uptake of embedded systems grows.

Given the possible electrification scenarios that may happen in the near future, this same analysis could have been developed but considering the roll-out of EV charging points, or CHP at the household level. It should be noted that the deployment of these technologies may have the similar effect as PVs; that is, they can lead to a higher level of imbalance in the network. As a consequence, it is very likely that in the future, there will be more factors that can trigger imbalances and losses. Therefore, it would be very important to study their impacts, and how they interact with one another. Even if these technologies are deployed, there are certain areas where it would be very difficult to have a 100% of penetration for SM, and therefore, methods such as the one put forward in this thesis will be required to keep the network as balanced as possible to avoid losses.

Regarding the limitations of the presented methods in this thesis, there are many areas of improvement as well:

- Further research on hierarchical clustering and its capacity to cluster data sets based on the correlation among voltage measurements is required. Since its low accuracy for lower SM penetration levels was attributed to the fact it works better for larger data sets. However, it would be interesting to investigate why voltage correlations are not enough for shaping clear clusters that allow phase identification.

- In addition, regarding the impact of PV-DG on the accuracy of the phase identification results obtained with k-means. While it is good news that moderate scenarios do not pose a problem for the accuracy of this method, it should be further investigated if higher levels of penetration will indeed affect the voltage

correlation values among customer's profiles. Thus, leading to inaccuracies in the phase matching of k-means clustering.

- In addition, from comparing the results from the three first circuits with the three last ones, it was concluded that average voltage measurements are more desirable since they present higher correlation values in SMs connected to the same phase. However, this is probably not enough evidence, and it should also be further investigated. This could be done by obtaining instantaneous and average voltage measurements for the same circuit, and comparing the results obtained with the k-means clustering. This way, more revealing results could be achieved. And therefore, network operators could decide on whether changing the current setting of SMs should be changed or not.

13 References

Arya, V. & Mitra, R., 2013. *Voltage-based clustering to identify connectivity relationships in distribution network.* Vancouver, Canada, IEE.

Arya, V. et al., 2014. *Voltage Analytics to Infer Customer Phase.* Istanbul, IEEE.

Arya, V. et al., 2011. Phase Identification in Smart Grids. *Architectures and Models for the Smart Grid (IEE SmartGridComm),* pp. 25-30.

Bandyopadhyay, S. et al., 2015. *Machine Learning for Inferring Phase Connectivity in Distribution Networks.* Miami, IEEE.

Begovic, M. M. et al., 2012. *Integration of Photovoltaic Distributed Generation in the Power Distribution.* Grand Wailea, Maui, IEE.

Berkhin, P., 2006. A Survey of Clustering Data Mining Techniques. In: J. Kogan, M. Teboulle & C. Nicholas, eds. *Grouping Multidimensional Data. Recent Advances in Clustering.* s.l.:Springer, pp. 25-71.

Blakely, L., Reno, M. J. & Feng, W.-c., 2019. *Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries.* Champaign, IL, USA, IEEE.

Brint, A., Poursharif, G., Black, M. & Marshall, M., 2018. Using Grouped Smart Meter Data in Phase Identification. *Computers and Operations Research,* Volumen 96, pp. 213-222.

Chen, C. S., Ku, T. T. & Lin, C. H., 2012. Design of Phase Identification System to Support Three-Phase Loading Balance of Distribution Feeders. *IEEE Transactions on Industry Applications,* 48(1), pp. 191-198.

Cipcigan, L. M. & Taylor, P. C., 2007. Investigation of the Reverse Power Flow Requirements of High Penetrations of Small-scale Embedded Generatoin. *IET Renewable Power Generation,* 1(3), pp. 160-166.

Cui, X. & Potok , T. E., 2005. Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm. *Journal of Computer Sciences,* Issue 5, pp. 27-33.

Cundeva, S., Bollen, M. & Schwanz, D., 2016. *Hosting Capacity of the Grid for Wind Generators Set by Voltage Magnitude Distortion Levels.* Belgrade, Serbia.

Dang-Ha, T.-H., Olsson, R. & Wang, H., 2015. *The Role of Big Data on Smart Grids.* Chengdu, IEEE, pp. 19-21.

Department for Business, Energy & Industrial Strategy, 2019. *National Statistic. Solar Photovoltaics Deployment.* [Online] Available at: https://www.gov.uk/government/statistics/solar-photovoltaics-deployment [Accessed August 3 2019].

Department for Business, Energy & Industrial Strategy, 2019. *Smart Meter Statistics. Quaterly Report to end March 2019.* [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/804767/2019_Q1_Smart_Meters_Report.pdf [Accessed 10 August 2019].

Department for Business, Energy & Industrial Strategy, 2019. *Statistical release and data: Smart Meters, Great Britain, quarter 1 2019.* [Online] Available at: https://www.gov.uk/government/statistics/statistical-release-and-data-smart-meters-great-britain-quarter-1-2019 [Accessed 10 August 2019].

Department of Energy & Climate Change, 2011. *Implementing the Climate Change Act 2008. The Government's proposal for setting the fourth carbon budget.* [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/48081/1683-4th-carbon-budget-policy-statement.pdf [Accessed 19 July 2019].

Dubey, A., Santoso, S. & Maitra, A., 2015. *Understanding Photovoltaic Hosting Capacity of Distribution Circuits.* Denver, CO, USA, IEE.

Fan, Z. et al., 2012. *The Power of Data: Data Analytics for M2M and Smart Grid.* Berlin, IEEE, pp. 1-8.

Gunasekaran, R. et al., 2017. Power Line Carrier Communication Using Automated Meter Reading. *Bioprocess Engineering*, 1(4), pp. 104-109.

Han, J., 2006. Foreword. In: J. Kogan, C. Nicholas & M. Teboulle, eds. *Grouping Multidimensional Data.Recent Advances in Clustering.* s.l.:Springer, pp. V-VI.

Hollingworth, D., Birch, A., Miller, D. & Lyons, P., 2012. *Demonstrating Enhanced Automatic Voltage Control for Today's Low Carbon Network.* Lisbon, Portugal, CIRED Workshop, pp. 1-4.

Hoornaert, F. et al., 2016. LV distribution network voltage control mechanism: Analysis of large-scale field-test. *Sustainable Energy, Grids and Networks,* Volume 6, pp. 7-13.

Jain, A. K., Murty, M. N. & Flynn, P. J., 1999. Data Clustering: A review. *ACM Computing Surveys,* 31(3), pp. 264-323.

Jayadev P, S., Rajeswaran, A., Bhatt, N. P. & Pasumarthy, R., 2016. *A Novel Approach for Phase Identification in Smart Grids Using Graph Theory and Principal Component Analysis.* Boston, MA, USA, IEEE.

Kai, W., Skiena, S. & Robertazzi, T. G., 2013. Phase balancing algorithms. *Electric Power System Research,* Volume 96, pp. 218-224.

Kong, W., Ma, K. & Wu, Q., 2018. Three-Power Imbalance Decomposition into Systematic Imbalance and Random Imbalance. *IEEE Transactions on Power Systems,* 33(3), pp. 3001-3012.

Li, Y. & Crossley, P. A., 2014. Voltage Balancing in Low-voltage radial feeders using Scott transformers. *IET Generation, Transmission & Distribution,* Volume 8, pp. 1489-1498.

Luan, W. y otros, 2015. Smart Meter Data Analytics for Distribution Network Connectivity Verification. *IEE Transactions on Smart Grid,* 6(4), pp. 1964-1971.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations. In: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probabilit.* Berkeley, California, University of California Press, pp. 281-297.

Ma, K., Li, R. & Li, F., 2015. Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance. *IEE Transactions on Power Systems,* 31(4), pp. 2885-2891.

Ma, K., Li, R. & Li, F., 2017. Utility-Scale Estimation of Additional Reinforcement COst From Three-Phase Imbalance Considering Thermal Constraints. *IEEE Transactions on Power Systems,* 32(5), pp. 3912-3923.

MathWorks, 2018. *GapEvaluation Criterion.* [En línea] Available at: https://es.mathworks.com/help/stats/clustering.evaluation.gapevaluation-class.html
[Último acceso: 11 August 2019].

MathWorks, 2018. *Hierarchical Clustering Documentation.* [Online] Available at: https://es.mathworks.com/help/stats/hierarchical-clustering.html?lang=en
[Accessed 5 August 2019].

MathWorks, 2018. *kmeans Documentation.* [Online] Available at: https://es.mathworks.com/help/stats/kmeans.html?s_tid=doc_ta#bues2hs
[Accessed August 5 2019].

MathWorks, 2019. *K-means clustering Documentation.* [Online] Available at: https://es.mathworks.com/help/stats/k-means-clustering.html
[Accessed 28 July 2019].

MathWorks, 2019. *plotconfusion Documentation.* [Online] Available at: https://es.mathworks.com/help/deeplearning/ref/plotconfusion.html?searchHighlight=confusion%20matrix&s_tid=doc_srchtitle
[Accessed 13 August 2019].

MathWorks, 2019. *Silhouette Documentation.* [Online] Available at: https://es.mathworks.com/help/stats/clustering.evaluation.silhouetteevaluation-class.html?s_tid=doc_ta
[Accessed 12 August 2019].

Mitra, R. et al., 2015. *Voltage Correlations in Smart Meter Data.* Sydney, ACM.

Mokhtar, M. et al., 2019. *Predicting the Voltage Distribution for Low Voltage Networks using Deep Learning.* Bucharest, IEE.

Morgan, G. M. et al., 2019. *The Many Meanings of "Smart Grids".* Pittsburgh: A Briefing Note from the Department of Engineering and Public Policy Carnegie Mellon University.

National Audit Office (NAO), 2018. *Rolling out of smart meters.* [Online] Available at: https://www.nao.org.uk/wp-content/uploads/2018/11/Rolling-out-smart-

meters.pdf
[Accessed 10 August 2019].

National Grid, 2018. *Future Energy Scenarios.* [Online] Available at: http://fes.nationalgrid.com/media/1363/fes-interactive-version-final.pdf [Accessed 30 July 2019].

Navarro, A., Ochoa, L. F. & Randles, D., 2013. *Monte Carlo-Bases Assessment of PV Impacts on Real UK Low Voltage Networks. In IEE IEEE Power and Energy Society General Meeting.* , IEEE Power Energy Soc. Gen. Meet.

Ochoa, L. F., Ciric, R. M., Padilha-Feltrin, A. & Harrison, G. P., 2005. *Evaluation of Distribution System Losses due to Load Unbalance.* Liege, Belgium, pp. 1-4.

Ofgem, 2010. *Electricity Distribution Units and Loss Percentages Summar.* [Online] Available at: https://www.ofgem.gov.uk/sites/default/files/docs/2010/08/distribution-units-and-loss-percentages-summary.pdf
[Accessed 7 August 2019].

Olivier, F., 2018. *Solutions for Integrating Photovoltaic Panels Into Low-voltage Distribution Networks,* Liège: Université de Liège.

Olivier, F., Ernst, D. & Fonteneau, R., 2017. *Automatic Phase Identification of Smart Meter Measurement Data.* Glasgow, CIRED.

Olivier, F. et al., 2018. *Phase Identification of Smart Meters By Clustering Voltage Measurements.* Dublin, Ireland, IEE, pp. 1-8.

Pappu, S. J., Bhatt, N., Pasumarthy, R. & Rajeswaran, A., 2018. Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data. *IEEE Transactions on Smart Grid,* 9(5), pp. 5113-5122.

Pezeshki, H. & Wolfs, P., 2012. *Correlation Based Method for Phase Identification in a Three Phase LV Distribution Network.* Bali, 22nd Australasian Universities Power Engineering Conference (AUPEC), pp. 1-7.

Pezeshki, H. & Wolfs, P. J., 2012. *Consumer Phase Identification in a Three Phase Unbalanced LV Distribution Network.* Berlin, 2012 IEEE.

Romero Agüero, J. & Steffel, S. J., 2011. *Integration Challenges of Photovoltaic Distributed Generation on Power Distribution Systems.* Detroit, MI, IEE.

Schwanz, D. et al., 2017. Stochastic Assessment of Voltage Unbalance Due to Single-Phase-Connected Solar Power. *IEEE Transactions on Power Delivery*, 32(2), pp. 852-861.

Seal, B. K. & McGranaghan, M. F., 2011. *Automatic Identification of Service Phase for Electricity Utility Customers.* Detroit, MI, USA, IEEE.

Short, T. A., 2013. Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling. *IEEE Transactions on Smart Grid*, 4(2), pp. 651-658.

Strbac, G. et al., 2014. *Management of electricity distribution network losses,* London: Imperial College / Sohn Associates.

Strbac, G. et al., 2018. *Strategies for Reducing Losses in Distribution Networks. Imperial College of London.* [Online]
Available at: https://www.ukpowernetworks.co.uk/losses/static/pdfs/strategies-for-reducing-losses-in-distribution-networks.d1b2a6f.pdf
[Accessed 29 July 2019].

Vathy-Fogarassy, Á. & Abonyi, J., 2013. Vector Quantisation and Topology Based Graph Representation. In: S. Zdonik, et al. eds. *Graph-Based Clustering and Data Visualization Algorithms.* London Heidelberg New York Dordrecht: Springer, pp. 1-16.

Wagstaff, K., Cardie, C., Rogers, S. & Schroedl, S., 2001. *Constrained K-means Clustering with Background Knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning.* San Francisco, Morgan Kaufmann Publishers Inc. , pp. 557-568.

Wang, W. et al., 2016. Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).*.

Watson, J. D., Welch, J. & Watson, N. R., 2016. Use of Smart-meter data to determine Distribution system topology. *IET The Journal of Engineering,* pp. 1-25.

Wen, M. H. F. et al., 2015. *Phase Identification in Distribution Network with Micro-Synchrophasors*. Denver, CO, USA, IEE.

Yan, R. & Saha, T. K., 2013. Investigation of Voltage Imbalance Due to Distribution Network Unbalanced Line Configurations and Load Levels. *IEE Transactions on Power Systems*, 28(2), pp. 1829-1838.

Zhu, J., Chow, M.-Y. & Zhang, F., 1998. Phase Balancing using Mixed-Integer Programming. *IEEE Transactions on Power Systems*, 13(4), pp. 1487-1492.