Data driven modelling of distribution system considering a high penetration of renewable energy sources for control applications

Author:
Carlo Viggiano

Supervisor:
Dr. Paul Trodden

This dissertation is submitted to the University of Sheffield in partial fulfilment for the degree of Doctor of Philosophy

Department of Automatic Control and Systems Engineering

Tuesday 1st August, 2023

To my family

Acknowledgements

First and foremost, I would like to thank God for granting me the necessary patience, strength, and support to complete this thesis. Nothing would have been possible without Him.

I would like to express my gratitude to my supervisor, Dr. Paul Trodden for his consistent advice and encouragement. His sincere and reflective guidance was fundamental in helping me shape and complete my research. I wish to express sincere gratitude to Prof. Eduardo Caicedo and Dr. Wilfredo Alfonso from Universidad del Valle, Colombia, with whom I developed a continuous collaboration that was invaluable to the completion of this thesis. I am also grateful with my sponsors for the financial support that made this PhD possible: the University of Sheffield and the Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS).

I greatly appreciate the assistance of all who contributed to the success of my programme, my peers, those within my research group and colleagues in the University. I will hold on to these lovely memories for the rest of my life.

I'm especially grateful to my colleague and mentor, Dr. Juan Gers, who introduced me formally to the wonderful world of power systems and modernisation of distribution systems, and who shares the passion for Smart Grids and Distribution Automation. Also, I would like to thank Dr. Anna Ferguson, Andy Holmes and Anthony Donoghue, who have given me the opportunity to continue working in this fascinating world in the UK.

Finally, I would like to thank my parents Martha and Tibaldo and my brother Julian, who have consistently supported me in all my projects throughtout my life and have encouraged me to give always my best. My deepest thanks are with family Gonzalez-Salas Duhne, especially Paulina. She has been a source of support and enthusiasm in the many hours devoted to the preparation of this thesis, and she will be the person to walk alongside me to overcome all new, promising and wonderful projects. Also, many thanks to all the friends I have around the world and the ones I have made in the UK, for all the support and wonderful memories that made the hardships more bearable.

Thank you all and God bless.

Abstract

The distribution system is undergoing a transition process of modernisation, where it is expected to make more efficient use of the available resources and equipment. This creates challenges for maintaining the system operating within the allowed operational conditions considering the constant changes of "bidirectional" power flows. To tackle these challenges, the use of available data may help to describe the distribution system without relaying on "snapshots" that represent worst-case scenarios or non-updated information from any of the traditional electric parameters used in modelling. This thesis focused on the definition of an approach to analyse and model the voltage in distribution systems with high penetration of renewable energy, and the (potential) use of these models on control applications. A review of the relevant data-driven modelling approaches was conducted, including the background of power systems parameters and modelling, voltage modelling and control for distribution systems with high penetration of renewables, and observability and controllability in distribution systems. Different modelling approaches in distribution systems were reviewed, to consider the potential inclusion of measurable data and new metrics to help detect the desired system dynamics to be represented. This data is a set of time-series measurements that are expected to describe the distribution system itself. An innovative description of the distribution system was introduced, by using a set of new proposed metrics. These metrics were based on power (dissipated) in lines provided and voltage covariance between nodes to describe, respectively, size and distance of perturbations. They give relevant spatial-temporal information of the distribution system and its perturbations based on time-series data. Different scenarios were explored to evaluate the limits of the amount of information that can be extracted from the distribution system. Results showed that the metrics can represent spatial-temporal features and events occurring in the distribution system, which make them suitable for real-time applications. Once the metrics were obtained, an algorithm was proposed to produce a linear model that represents the distribution system using measurable data. This algorithm requires a revision of data to understand the structure of the proposed model. In this case, the time-series data must be stationary to produce the linear model. Once this condition was achieved, the next step was to produce a provisional model to explore the proposed regressors used in modelling. Once a reference case was obtained, the input data was analysed to detect any conditions that may introduce errors in the model (e.g., collinearity in some exogenous regressors), improve the response (e.g., analysis of cross-validation using Granger-causality) and highlight the regressors and

time lags that improve the response. The first and final models were compared to explore if the proposed metrics could explain the system dynamics. Results after applying the algorithm showed that it was possible to obtain a good model for one-step ahead prediction, which can be easily integrated to any control structure. Finally, the refined model was improved by presenting a prediction interval based on bootstrapping and cross-validation techniques used in time-series data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis aims to define an approach that assists in the analysis and modelling of the voltage in distribution systems with high penetration of renewable energy sources, primarily for controlling applications. Distribution systems have become interesting connection points for the electric network to obtain energy. However, grid operators must consider making changes in their planning, operation, and control to successfully integrate these resources into the power system. Distribution systems were not designed as points for injecting energy into the network, and rebuilding the entire power system to achieve this goal is not feasible. That is the main reason to understand, model and control the grid in a different way, avoiding regular practices such as curtailing renewable energy in several scenarios. This chapter sets the stage for the work carried out in this thesis with an overview of the transition of the distribution system, motivation and related aims linked to data-driven characterisation and modelling, the path for a possible application of model predictive control approach in voltage regulation and a description of the scope and objectives of this thesis, laying the groundwork for the work carried out in this thesis.

1.1 Voltage Control in Modern Distribution Systems

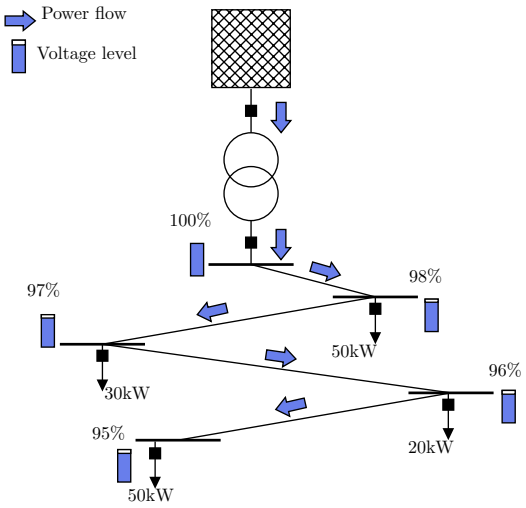
The increasing interest of governments in reducing CO_2 emissions has heightened the need to improve the distribution system to allow for high penetration of renewable energy sources. Distribution systems were built to play a different role in the power system than what is required today. Their integration into the system is achieved at different levels, from concentrated big generation parks to remote small units close to the load [1]. This modernisation is expected to be achieved without high additional investment in the power grid infrastructure, which means

efficient use of resources and equipment [2]. New operational conditions and functionalities are expected under this scenario, including bidirectional power flows. This change not only reduces fossil fuel-based generation but also reduces electrical losses around the system due to the transportation of energy [3, 4]. A key challenge that emerges under this new regime of operation is maintaining the system voltage within acceptable limits, despite the high variability and uncertainty associated with renewable energy sources [5].

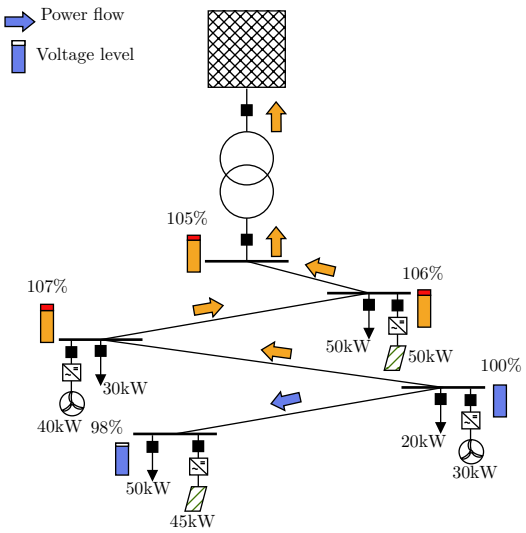
Figure 1.1 shows an image of the modern distribution system and the assets expected to be controlled. Figure 1.1a illustrates the distribution system under real conditions, including uncertainties associated with generation units (i.e., location of generation units, rated power, and power availability). The first two presented scenarios are considered fixed in planning studies, as these are usually known and determined before they are connected to the grid. However, for operational studies, it is unpredictable *where* and *which size* of new generation units will be connected. In a traditional radial distribution system with unidirectional power flow, the voltage drops when it is getting far from the feeder, and the network is operated and controlled under these assumptions (the load tap changers, voltage regulators, and capacitor banks are adjusted according to these operational considerations). Current conditions of distribution systems allow the integration of renewables without violating the operational constraints when the generated energy is all consumed close to the load. The main issue in this integration is illustrated in Figure 1.1b. Power injected by the renewable generator units can exceed the consumption of the closest loads. Therefore, the power will flow back to the feeder, which can increase the voltage and become an operational issue [6, 7]. The traditional problem for a distribution system is maintaining the voltage over the lower limit because voltage drops in the distribution lines. For a distribution system with renewable energy technologies, the problem is the opposite because the voltage tends to increase above the upper limit when there is reverse power flow.

1.2 Motivation

Existing solutions for reducing overvoltage typically comprise a blend of local automatic controls (e.g., voltage regulators associated with an asset) and global (network-level) decisions (e.g., curtailment of renewable energy injection; feeder voltage change). The limitation of the former is the lack of coordination among actions, while in the latter control decisions are typically made according to a static network analysis and/or trial and error rules. This is the most common solution to overcome the increase in voltage, as illustrated in Figure 1.2. However, this solu-



(a) Distribution system with no operational issues



(b) Distribution system with voltage issues

Figure 1.1: Example of a modern distribution system and its assets with integration of renewable energy under different operational conditions

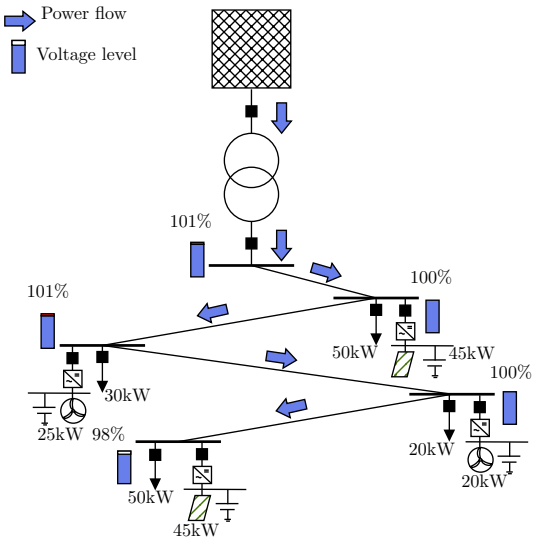


Figure 1.2: Example of an expected operational outcome for distribution systems with integration of renewable energy when correct control scheme is applied

tion reduces the amount of renewable energy that can be integrated into the power system, which contradicts the desire to use as much renewable energy as possible. It also raises interesting but challenging questions regarding the compensation for prosumers.

According to what is mentioned before, the control approach aims to be dynamic, adaptive, robust, and scalable. Why? The distribution network is a dy-

namic system, and proper control design requires a dynamic model; if an acceptably accurate model can be obtained, it is anticipated that the performance of the model-based controlled distribution system would exceed what is possible via the state-of-the-art approaches based on static analyses and uncoordinated local controllers. The model-based control approach should be adaptive in the sense that it is capable of self-tuning or self-reconfiguring following the network's actual status, such as detecting system topology and faults in distribution systems or accessing real-time measurements of power consumption/generation profiles. It should be robust to the uncertainties that pervade the distribution network. Finally, it must be scalable to manage a growing amount of data, monitoring, and control decisions by adding components to the system.

Figure 1.3 presents the control idea proposed in the IEEE in Smart Grid environment vision [8], based on the scenario in which the new control concept previously presented is adopted. It is important to highlight the construction of the distribution system model based on data in this control approach. This model enhances the definition of voltage set points, taking into consideration operational constraints. Also, variability of renewable energy sources is expected to be an input instead of a disturbance, as done in classical control approaches. This can give more flexibility in the control targets and relax the limit restrictions (especially for renewable energy generation units integrated in the system). The main goal for this control problem is to improve the amount of renewable energy sources without affecting voltage levels, which might be pursued via different devices such as OLTC, voltage regulator, batteries, D-STATCOMs, among others.

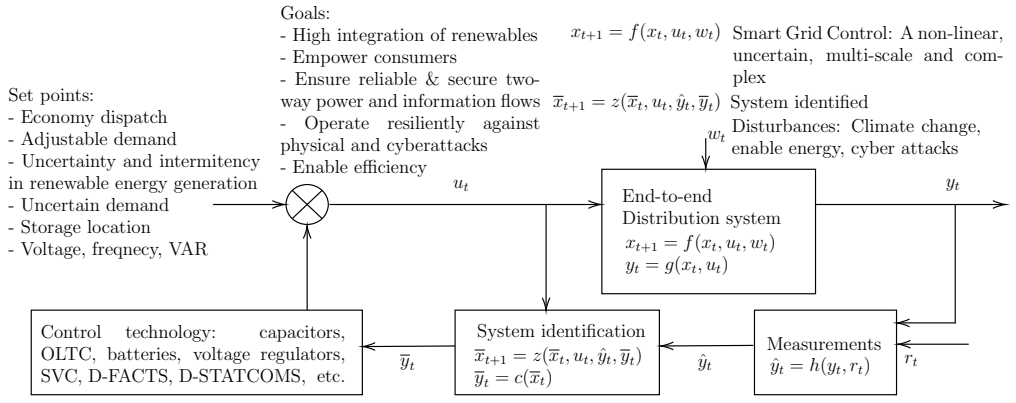


Figure 1.3: Proposed modern control scheme for distribution systems in Smart Grid environments based on modernisation vision introduced in [8]

The previous idea can be summarised in the classic controllability and observability problem, which has been discussed in the context of power systems [9–13]. They have exposed the requirement of obtaining a model that contains

relevant information on internal states and can be changed by changing the system inputs (e.g., power injections). Additionally, the non-measured states should be estimated by measuring the relationship input-output. However, this idea of controllability and observability is focused on maintaining the classical stability of the power system (rotor angle of generators, voltage, and frequency). This term of "dynamic" is slightly different from the traditional dynamic system known in power systems, and it will not be limited only to the action of traditional assets (e.g., voltage regulators or inverter controllers). It is expected to capture the long-term effects of loads and renewable inputs on key network voltages as the output to be controlled (e.g., load profiles according to the time of the year, irradiance levels in different seasons, among others). If something different is not indicated, the long-term dynamics on which this thesis is focused are referred with the term "quasi-dynamics". This requires the execution of multiple load flow calculations performed at discrete time instances. The time instances are defined by the user, and in this thesis, they are determined to capture the change in system voltages and power flows caused by changes in supply and demand on the scale of seconds to minutes. When the system is partially observable, it is required to analyse all the available data taken from measurements to select which information is providing the main picture that can describe the voltage profile and help in tracking to the desired state. The identification of these quasi-dynamics could be of valuable use in developing voltage control approaches for distribution systems with high penetration of renewables yet low availability of measurements. This is the motivation for the present thesis.

The primary challenge is the same one encountered whenever a new controller for any system is to be designed: modelling. Typically, this is done using a physics-based structure, characterised by the topology and parameters of the network. Physics-based models are easy to develop for electricity networks and have been useful and successful in static analysis (e.g., state estimation; load flow), low-level control (e.g., of inverters), high-level control (e.g., Automatic Generation Control (AGC)), and operational decision making (e.g., dispatch and planning). They are used in Active Network Management (ANM) schemes, which has emerged as a technology to provide real-time monitoring and control in future distribution systems [14]. System model becomes the heart of this approach, which operate in conjunction with real-time measurements to capture and forecast system behaviour, detect significant changes to system variables, and provide appropriate control actions or coordinated signals. However, building such a model is particularly challenging for legacy distribution systems because (i) physics-based models scale linearly with the number of nodes in the network, which may run into the

thousands, leading to a large-scale model that is excessively large compared to the scale of the problem (i.e., typically, over-voltage occurs at a selective subset of "bottlenecks" nodes) (ii) the voltage quasi-dynamics of the distribution network are primarily a consequence of the load, which typically comprises a broad mix of different devices and systems, thus is hard to model from first principles, and (iii) in practice, operators may not have accurate or complete information on network assets and system parameters.

For these reasons, the current thesis aims to adopt a data-driven approach to obtain a model that represents the distribution system and is suitable for control applications. Approaches based on system identification techniques have seen great success in traditional industrial control for precisely these reasons: they can accurately describe, with low-order models, complex processes that are difficult or undesirable to model from first principles [15–18]. Moreover, system identification modelling goes hand in hand with adaptive control: system models are easily refined and updated on-line, using the most recent system data, whereas physical modelling requires either knowledge of the change or re-modelling from scratch [19, 20]. Nevertheless, obtaining these models is not just an application of off-the-shelf system identification methods to a distribution system dataset. There is an important challenge first in identifying and selecting relevant inputs and outputs that can be measurable and suitable to capture the relevant quasi-dynamics required in the model. This challenge is exacerbated in distribution systems because in legacy systems, knowledge of system topology and parameters is poor.

1.3 Definitions, terminology and general considerations

In this section, relevant definitions, terminology and general considerations are specified that are used throughout this thesis. Unless stated otherwise, the following terms and assumptions should be considered as given:

1.3.1 Definitions and terminologies

- System dynamics [21] corresponds to a methodology and mathematical modelling technique used to analyse and study the dynamic behaviour of complex systems over time. To achieve this, it is required to recognise the relationships and interactions among the components of a system, including feedback loops and time delays. In a system dynamics approach, a system is viewed as one in which its state changes continuously over time. A "state" refers to a set of variables that collectively characterise the current condition of the system and its components. These state variables are used to predict

the future behaviour of the system and how it responds to different inputs and external influences.

- Dynamics [22] refers to the changes over time of the states of a system in response to inputs or disturbances. It involves the study of the time evolution of system variables or states and how they interact with each other to produce the system's behaviour. In the context of power systems, these dynamic processes include the electrical machines and inverter-based generation , system generation governing and prime-mover energy supply, which are analysed in different time ranges that vary from microsecond to minutes. The time-range classification is relevant to set the component modelling.

- Quasi-dynamics [23, 24] correspond to the long-term dynamics that represent the changes of variables, including load consumption, generation profiles, and exogenous variables of power systems such as solar irradiance. Even if they are represented by several load flow calculations (which are steady-state calculations), the time-series profile of each component produces variables that change "slowly" over time.

- Dynamic model [21] is a simplified representation of real-world entities in equations that mimics essential features of the system under study. Since the model is "dynamic", then its properties change over time. The construction of these models is constrained by the variables that can be measured, either to estimate parameters that are part of the model formulation or to validate model predictions.

- Voltage dynamics [25] are those dynamics that capture the changes of voltages for different time scales.

- Stability [26] in power system context is defined as the property of a power system that enables it to remain in a state of operating equilibrium under normal operating conditions and to regain an acceptable state of equilibrium after being subjected to a disturbance.

- Voltage stability [27–29] is the ability of a power system to maintain steady acceptable voltages at all buses in the system under normal operating conditions and after being subjected to a disturbance:

- Linearization [21] is a method for assessing the local stability of an equilibrium point of a system of nonlinear differential equations or discrete dynamical systems.

- Critical perturbation corresponds to a change in the consumption or generation that brings the operational values (voltages) outside expected boundaries (which normally corresponds to operational limits).

- Loadability [2] corresponds to the maximum load that a distribution system component can handle before encountering an electrical or operational constraint limitation.

- Measurable data [30] is a broader term that refers to any information or variables that can be quantified or observed using measurement devices. This is typically in raw form and may not have a specific context or interpretation until it is processed or analysed.

- Measurements [30] refer to the process of obtaining quantitative values or observations of specific properties or variables using instruments or techniques. Measurements involve the application of standard units and scales to quantify physical characteristics, quantities, or attributes. The data obtained through measurements are the result of this process and represent the specific values or observations recorded at a particular time or location. These include voltage and power magnitudes.

- Adaptive [31] refers to the property of a system, component, or controller of being able to change certain characteristics or behaviour according to new circumstances or changes in the environment, in order to adapt to new conditions without requiring prior information about the bounds on uncertain or time-varying parameters.

- Robust [29, 31] refers to the property of a system, component, or controller of being able to deal with uncertainty, for which some a priori information, such as bounded modelling errors, is required. Therefore, its parameters are fixed based on this knowledge.

- Scalable [21] refers to the property of a control system or controller that can be adapted or expanded to handle larger or more complex tasks without significant changes to its fundamental structure, while still maintaining its performance and functionality, and without losing efficiency or effectiveness.

- Controllability [21] refers to the property of a dynamical system of being driven from any initial state to any desired final state within a certain time frame (or being able to reach any state within its state space) by applying suitable control signals or appropriate control inputs.

- Observability [21] refers to the property of a dynamical system that allows its internal states to be inferred or estimated from the available measurements of its outputs over a certain time period. Therefore, an observable system is one in which all the internal states can be reconstructed or observed based on the available output measurements and the knowledge of the system dynamics.

1.3.2 Devices and control schemes

Some of the common devices used in distribution systems for voltage control are voltage regulators [32], which are auto transformers capable of increasing or reducing voltage to maintain system voltage levels within required ranges by sensing system voltage and adjusting their tap changers. Additionally, the transformer at the main feeder has its own OLTC to regulate the main substation node voltage. Around the system, there are capacitor banks, which are systems consisting of several capacitors connected in series or parallel to form an energy storage system and change voltage levels by injecting reactive power. In a similar way, inverter-based technologies, including Photovoltaic (PV) units units and batteries, inject active power that can affect the voltage, especially in distribution systems with a high resistive component in the topology.

The control schemes commonly used for voltage control using the mentioned devices can be summarised in conventional control architectures (local or remote), in which the voltage setpoints can be defined locally (by sensing the voltage at the same point where the device is installed), or remotely by solving load flows that reflect the best operational condition to achieve the desired voltage level (usually at the main substation feeder). Voltage/VAR optimisation is another approach, where an SCADA system mainly defines the action of the devices to achieve the desired voltage objective. Nowadays, there are new trends of using these devices to control voltage, such as voltage control by using PV units, decentralised voltage control, and the combination of these with Volt-Var optimisation. As an example and illustrated in Figure 1.1, the system has none of the components mentioned above to regulate the voltages over the nodes. The orchestrated action of batteries and OLTCs considerably improves the voltage over nodes, as shown in Figure 1.2.

For this thesis, only the action of voltage regulators and capacitor banks is considered, since they are part of the distribution system used as a reference. Appendix E presents more details about the components of the reference case. The control action predefined for this network used as reference corresponds to fixed settings installed devices and for some nodes across the system (local approach). Nevertheless, the procedure evaluated in this thesis should not be limited to the

strategy defined, as it is only reflected in the final model obtained independently of the control approach.

1.3.3 Considerations for reference data and measured data

The data used in this thesis was synthetically produced according to the procedure presented in Section 3.3, with a resolution of 1 minute. The system used as a reference is presented in Appendix E. The measurements used were a sampling of this reference data every 10 minutes, as indicated in Chapters 3 and 4. The sampling rate of 10 minutes is a realistic window for several real applications that meet the requirement to properly describe the slow dynamics to be characterised in this model approach. For this thesis, it is assumed that the data has no noise, which can considerably affect the results in real applications. Filtering and processing techniques should be applied in the presence of noise. However, as the main purpose of the thesis is to develop data-driven modelling based purely on data processing, the main focus of the approach is to understand what information can be extracted from the measurable data to obtain relevant characteristics of the system that can describe what is exactly happening within the system in a specific window of time. Therefore, it is not required to focus on the filtering and processing of noisy data, which is a topic highly covered in other theses [33–35].

1.4 Thesis scope

This thesis aims to address the goals presented in Section 1.2. Specifically, this study endeavours to propose an approach for analysing and modelling distribution systems with high levels of renewable energy and considering the associated uncertainties and operational constraints. The obtained models can be used for voltage prediction and potentially for voltage control. Before attempting any modelling and control efforts, it is required to comprehend three essential elements: (i) what to measure – how informative is it about the required variables to be controlled? (ii) what to change/manipulate – how powerful is it with respect to managing these variables? (iii) what the fundamental cause-effect relationship is between what is changed and what is measured and, ultimately, the components that are required to be controlled. A key part of the aim is understanding these aspects first, and developing an approach to solving the problem, with as little prior knowledge and assumptions as possible. This major task can be divided into the following smaller objectives:

1. To review and highlight the range of data-driven modelling approaches and their applications to control voltages in conventional and future distribution grids: before starting any regressor selection process or data-modelling approach, it is required to review the state-of-art for different techniques currently used in distribution system applications and understand their limitations of representing actual dynamics of the network. Most of these approaches are physics-parameter-based, which can limit their scalability when it is required to model a system with several nodes. Additionally, it is challenging and expensive for some distribution systems to get access to all required data or measurements to build these models. Therefore, different modelling approaches and potential strategies to overcome these limitations will be presented, by using new metrics, which can be obtained from measured time-series data and represents the system quasi-dynamics.

2. To identify the critical factors, metrics and time-series measurements that helps to describe relevant features of the distribution systems for maintaining the voltage in acceptable operational ranges: once the required characteristics to overcome traditional modelling approaches are discussed, there is required to review the features of distribution systems that are measurable and helps on sketch a representation of the required quasi-dynamics. For instance, spatial characteristics such as electric distance or evolution of voltage covariance in time, are suitable to describe the system without accessing all internal states. The scope and limitations of these metrics to catch the voltage quasi-dynamics will be part of the discussion to fulfil this objective.

3. To develop and compare different data-driven modelling approaches for voltage quasi-dynamics in distribution networks, resulting in reduced-order, accurate models for control: after the discussion and selection of most representative metrics that helps to describe the distribution system, the next step corresponds to apply different regression techniques that are suitable for time-series measurements. It is desired to obtain linear state-space representations that are easily to scale and control. Therefore, autoregressive regressions or Koopman-based techniques are good candidates to fulfil this requirement. The application of these methods and the discussion of relevant feature of each technique will be part of this objective.

4. To propose an algorithm that estimates actual condition of the distribution system using available measurements and build a model based on data: the final objective for this thesis is to obtain a systematic procedure that may be potentially used in real-time applications for modelling (and controlling) the

distribution system. Therefore, statistical analysis of the data and the models are expected to be used and enhance the interpretation of the obtained results. Additionally, it can give insights on the procedure to increase the performance of model results based on the analysis of their residuals and the validation of statistical assumptions. Additionally, the incorporation of features such as prediction intervals may help to explain the boundaries of system responses.

1.5 List of Publications

Some of the work presented in this thesis has also been published and/or prepared for submission:

1.5.1 Published Papers

1. Carlo Viggiano, Paul Trodden, Eduardo Caicedo and Wilfredo Alfonso, "Data-Driven Characterisation of Distribution Systems for Modelling and Control Applications," 2022 International Conference on Smart Energy Systems and Technologies (SEST), Eindhoven, Netherlands, 2022, pp. 1-6, doi: 10.1109 / SEST53650.2022.989842. (Appendix A).

1.5.2 Submitted Papers and Papers Under Preparation

1. Carlo Viggiano, Paul Trodden, Eduardo Caicedo and Wilfredo Alfonso, "Distribution Systems Modelling by Data-Driven Voltage Characterisation for Control Applications-Part I: Input Analysis", journal paper 2022. *Submitted* (Appendix B).

2. Carlo Viggiano, Paul Trodden, Eduardo Caicedo and Wilfredo Alfonso, "Distribution Systems Modelling by Data-Driven Voltage Characterisation for Control Applications-Part II: Case Studies", journal paper 2022. *Submitted* (Appendix C).

3. Carlo Viggiano, Paul Trodden, Eduardo Caicedo and Wilfredo Alfonso, "Data-Driven Time-Series-based approach for modelling of distribution system with high penetration of renewable energy sources", journal paper 2023. *Under preparation* (Appendix D).

1.6 Thesis Outline and Main Thesis Contributions

Chapters outline and main contributions of this thesis are presented as follow:

Chapter 2 provides answers to objective 1 by reviewing and highlighting the range of data-driven modelling approaches. The literature review in this chapter focuses on various topics that are highly relevant to the aims of the research. These include power systems parameters and modelling, voltage modelling and control for distribution systems with high penetration of renewables, observability and controllability in distribution systems, and data-driven modelling of time-series background. The basics of power flow are covered in the first section of this chapter. The power-voltage equation is presented for non-lossless systems, and the discussion regarding the scenarios of voltage operational restriction is presented to define the study regime for distribution systems. The possible control objectives for voltage in distribution systems are presented, and the state-of-the-art approaches are discussed to highlight the advantages and opportunities for research to increase the possibilities of voltage control in distribution systems. Following this, the controllability and observability problems in distribution systems are discussed. The chapter concludes by presenting the requirement of developing data-driven models using measurable data. System identification is presented as a tool for modelling in modern distribution systems, as well as different potential techniques to produce state-space representations based on data. Finally, various uncertainty analysis approaches and their relevance to analysing renewable energy sources are discussed.

Chapter 3 provides answers to objective 2, which identifies the critical factors, metrics, and time-series measurements that help describe relevant features of distribution systems. Chapter 2 showed the feasibility of using data to produce models of the distribution system for voltage control application. Chapter 3 offers a novel description of the distribution system using measurable data that is useful in the modelling process. As a result, one of the original contributions of this chapter is a data-driven approach to characterising distribution systems based only on time-series measurements, which includes relevant spatial-temporal information of the distribution system and its perturbations. An analysis of the power injections was conducted to understand the impact of active and reactive power on voltage variation. The active power (dissipated) and the reactive power (stored) in lines provided useful information about the system topology and the relevance of each type of power in the potential control action, without any other previous information, such as electric line parameters, nodes connectivity, etc. This description can be used as potential inputs in a control application. A matrix of

measured power between nodes was enhanced using voltage correlations, which is the proposed metric to describe the size of the power perturbation. In addition, the voltage covariance was studied to describe information on nodes connectivity and electrical distance. The average value of 'normalised' voltage covariance was proposed as a potential metric to describe the distance of perturbation action. Therefore, another point that made this thesis different from related ones was the definition of a methodology to analyse and reduce model complexity based on electrical distance by using Fisher z-transform and voltage covariance from measured data. Different scenarios and the number of measurable nodes were evaluated to explore the limits of the information that can be extracted from the proposed metrics, which was also part of the contributions in this chapter. Both metrics showed high potential to be used in the process of building models for voltage control applications. A final contribution in this chapter was an interpretation of the new metrics and their potential application in control applications. The work presented in this chapter was used to produce a conference paper [36] and a two-part journal paper [37, 38].

Chapter 4 presents results that provide answers for objective 3, which is to develop and compare different data-driven reduced-order modelling approaches, and objective 4, which is to propose an algorithm that estimates the actual condition of the distribution system using available data. In this chapter, the use of metrics proposed in the previous chapter to obtain a data-driven time-series modelling approach was introduced. A revision of the required conditions over data was done to produce linear models, which involved checking the stationarity of the data. Additionally, an evaluation of initial assumptions was conducted to determine if the initial model guessing was adequate to explain the voltage quasi-dynamics. The dataset was also revised to select critical scenarios that could represent the desired voltage quasi-dynamics to be modelled. One of the main contributions in this chapter was the implementation of an analysis of data distribution shape, collinearity, and analysis of cross-validation using Granger-causality concept to reduce the regressors and lags into only the selected relevant variables that improve the performance of linear regression. Different structure and model regression techniques were presented and discussed for this specific problem. Another analysis of new residuals was conducted and compared with the first attempt, to understand the impact of the selection of variables' methodology. As a result, one of the significant contributions in this chapter was a proposed data-driven approach to obtain a reduced-order linear representation of the distribution systems that consider exogenous variables. Finally, the integration of prediction interval based on bootstrapping and cross-validation techniques was

explored and discussed. Results showed the impact of the selected variables and the statistical validity of the obtained models. Additionally, a methodology was defined to review and improve the performance of models obtained by verifying initial statistical assumptions, which provides an insight into how they can be improved. This was a significant difference from this thesis with respect to similar ones. This work was used to produce a journal paper [39].

Chapter 5 is for concluding statements and contributions. Future research directions are also presented providing insights into some of the opportunities for building upon the work done in this thesis.

Chapter 2

Background and literature review

Previous chapter presented the requirements to achieve a modelling approach that takes advantage of data to produce models in modern distribution systems. Also, the aims of this thesis were introduced, and it is required to review and highlight a range of data-driven modelling approaches available from the state-of-the-art, especially evaluate their potential to control voltages in distribution grids. To achieve this goal, the first task is understanding the regime in which the voltage in distribution behaves in presence of stochasticity associated with consumers behaviours, renewable energy location and injection, among other exogenous variables that require to be explored. This would help to understand the limitations of the current modelling approaches to deal with random variables in presence of no previous knowledge of the distribution system. Therefore, it is introduced in Section 2.1 a background of distribution system modelling within the concept of voltage stability. Since the spectrum of the voltage regime under stability concept is broad, a revision of the general expressions is presented in Section 2.2 to narrow the problem to a specific context based on the general condition expected in distribution systems. Once the context of the problem is delineated and profiled for distribution system applications, Section 2.3 presents the direction and limitations of the state-of-the-art to deal with the voltage control problem under the current modelling approach in the assumptions presented in the formulation of this thesis. This helps to conclude that another perspective is required to introduce a solution that tackles this limitation. Therefore, a presentation of controllability and observability is introduced in Section 2.4 as an alternative to deal with this problem. The notion of considering the distribution system as a plant in relation to its controllability and observability is not novel; however, it has not been previously linked in literature as a means to address the modelling of voltage quasi-dynamics that encompasses exogenous random variables. This constitutes

a key contribution of the present thesis, and it is discussed in section 4.4 and particularly in section 4.4.2. From this section, it is introduced the relationship of how this information explains the internal dynamics that were discussed in the original distribution system background. Finally, the available time-series data in distribution systems and the limitations of current metrics to feed the modelling approach (which is related in most cases with system identification approaches) is presented in Section 2.5.

2.1 Understanding basics on the voltage stability problem in distribution systems

The main objective of this thesis is to develop models that can be used to control voltage operation in distribution systems, taking into account the high penetration of renewable energy resources. Therefore, it is necessary to discuss and analyse voltage dynamic concepts and operational voltage issues in environments with high renewable energy penetration. This introduction is intended only as a preliminary exploration to understand the nature of the relationship between voltage and power based on the complete mathematical expression. The obtained expression will also provide an idea of how the X/R ratio can affect the evolution of voltage when the power value changes. This will help to generate preliminary ideas about the shape of the proposed model, which will guide the discussion in Chapter 3, and more specifically in Section 3.4. This idea is intended to be implemented for both radial and meshed distribution systems, with the latter being difficult to evaluate in terms of voltage profiles.

Traditionally, voltage dynamics in power systems is related with voltage stability, which is a relevant operational concern [40]. Historically, there have been several discussions regarding the definition of stability (for rotor angle, frequency and voltage stability) [41, 42]. According to Kundur et al. [26], voltage stability refers to the system capacity to maintain voltages at acceptable levels after a disturbance from a given initial operating condition. This depends on power system ability to maintain the equilibrium between demand and supply. Instability incurs in a progressive fall or increase of voltage in some nodes. Sometimes the term voltage collapse is used, which is a sequence of events that brings voltage instability and finally blackout or low voltages.

The voltage stability analysis can be divided into small-disturbance and large-disturbance voltage stability [26]. On the one hand, small-disturbance voltage stability evaluates the ability of the electrical system to maintain voltages at accepted levels when there are small perturbations. This analysis studies the influence

of continuous controls, discrete controls, and the characteristics of loads at a specific instant in time. On the other hand, large-disturbance voltage stability refers to the analysis of the system's ability to maintain steady voltages during large disturbances such as circuit contingencies, system faults, or loss of generation. For generation loss, it is necessary to examine the nonlinear response of the power system over an appropriate period, including the interaction and performance of devices such as motors and transformer tap changers. The study period of interest may extend from a few seconds to tens of minutes.

Considering the previous classification, voltage collapses can be analysed according to the time scales that the event occurs as follows [43]:

1. In the range of seconds, electromechanical transients (e.g., generators, induction machines, DC components of short circuit currents) and power electronics (e.g., SVC, D-STATCOM).

2. In the range of tens of seconds, discrete switching devices (e.g., load tap-changers, excitation limiters)

3. In the range of several minutes, load recovery processes.

According to the list mentioned above, the first-time scale is called the transient time scale. The second and the third time scale correspond to the 'long-term' time scale. Figure 2.1 outlines a power system model relevant to voltage analysis as was stated above.

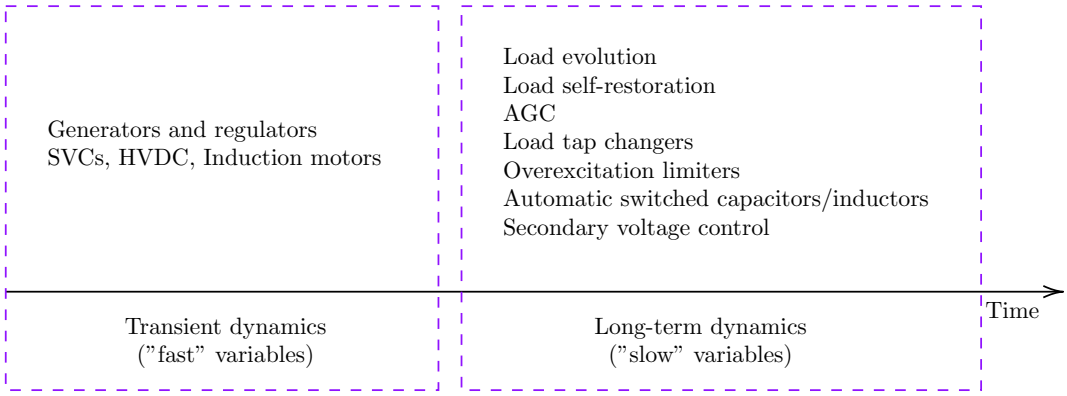


Figure 2.1: Time scales for voltage control

The power system model must be analysed to understand the voltage instability condition. Power systems are modelled based on non-linear differential algebraic equations, which arise as a consequence of an electric network of nodes where power is produced or consumed, interconnected by lines where power flows according to physics. This is represented by equation (2.1):

$$\dot{\mathbf{x}} = f(\mathbf{x}, \lambda) \tag{2.1}$$

where:

- $\mathbf{x} \in R^n$ represents a state vector, including the node voltage magnitudes (V) and angles (δ).
- $\lambda \in R^m$ is a parameter vector that represents the real and reactive power demand or supplied at each node.

The stable operational point corresponds to the conditions in which the power system's power flow is represented as shown in equation (2.2):

$$\mathbf{0} = f(\mathbf{x}, \lambda) \tag{2.2}$$

Considering these terms, instability is a condition in which there is no long-term equilibrium. More precisely, a solution of (2.2) could be either a stable equilibrium point or an unstable equilibrium point. The former has the property that for any chosen epsilon $\epsilon_s > 0$, there exists a delta $\delta_s > 0$ such that

$$||x(0) - x_{unstable}|| < \delta_s \Rightarrow ||x(k) - x_{unstable}|| < \epsilon_s \tag{2.3}$$

for all $k > 0$. The unstable point does not have this property, and the system states tend to move away from it under the smallest perturbations. For instance, during a restoration process, if power loads surpass the connected generation capability considerably, or when a post-disturbance steady-state operating point is small-disturbance unstable, or when there is a lack of attraction toward the stable post-disturbance equilibrium.

2.2 Voltage expression for distribution systems with non-lossless condition

To understand the behaviour of the system introduced in Equation (2.2), it is assumed that the parameter λ varies quasistatically in time, which makes the system with time-varying λ well approximated by keeping λ constant while the quasi-dynamics of the system act [25]. As it is assumed that the active and reactive power models are constant, from a quasi-dynamics perspective, power will remain the same between periods of time when the measurements are taken (more than one minute between two points). To illustrate voltage changes according to this quasi-dynamic expression, this assumption is regularly made to analyse the

problem from an "operating point", which is a stable equilibrium point and can be obtained from a reduced representation of the power system [44].

An intuitive way to understand the complexity of the problem is to analyse the voltage expression when power is transferred from one node to another. Therefore, the analysis begins with the equation development presented in [45], where any node can be approximately reduced to an equivalent node, allowing for a simplified analysis of the relationship between voltage and power, as illustrated in Figure 2.2. As previously mentioned, the analysis assumes constant power load type, meaning that the load at the receiving end consumes constant power regardless of changes in voltage. The power S_s is transferred from upstream node s with voltage $V_s\angle\varphi_s$ to downstream node r with voltage $V_r\angle\varphi_r$ and associated current $I_r\angle\alpha_r$. Complex values are assumed, including for the equivalent impedance between two nodes since the X/R ratio is unknown [45, 46].

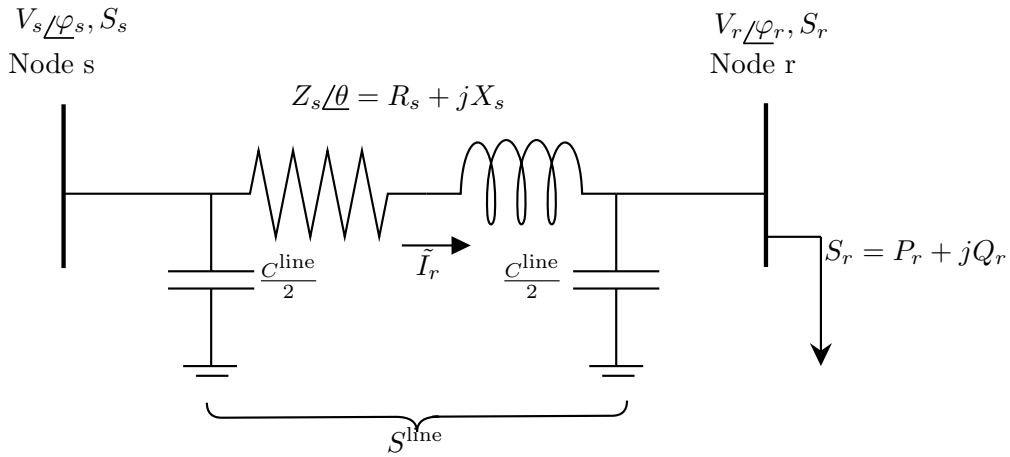


Figure 2.2: Power transfer between two nodes of the distribution system

The voltage at node r is

$$V_r\angle\varphi_r = V_s\angle\varphi_s - Z_s\angle\theta I_r\angle\alpha_r, \tag{2.4}$$

where $Z_s\angle\theta$ (also represented in its complex form $Z_s\angle\theta = R_s + jX_s$) is the line equivalent impedance with an angle θ seen from the beginning of feeder.

S_r represents the transferred power from node s to node r, which supplies the load (represented in its complex form as $S_r = P_r + jQ_r$). According to the Figure 2.2, an additional effect is considered because of the nodes' capacitance. This power

can also be expressed also according to the following expressions:

$$\begin{aligned} S_r &= V_r \angle \varphi_r [I_r \angle \alpha_r]^* + Q^{\text{line}} \angle 90^\circ \\ &= V_r \angle \varphi_r \left[\frac{V_s \angle (\varphi_s) - V_r \angle (\varphi_r)}{Z_s \angle \theta} \right]^* + Q^{\text{line}} \angle 90^\circ \\ &= V_r \angle \varphi_r \left[\frac{(V_s \angle (- \varphi_s) - V_r \angle (- \varphi_r)}{Z_s \angle -\theta} \right] + Q^{\text{line}} \angle 90^\circ \\ &= \frac{V_s V_r}{Z_s} \angle (\varphi_r - \varphi_s + \theta) - \frac{V_r^2}{Z_s} \angle \theta + Q^{\text{line}} \angle 90^\circ \\ &= \frac{V_s V_r}{Z_s} \angle (\delta + \theta) - \frac{V_r^2}{Z_s} \angle \theta + Q^{\text{line}} \angle 90^\circ \end{aligned} \tag{2.5}$$

where the angle difference between both nodes is $\delta = \varphi_r - \varphi_s$. From the last expression, apparent power can be split into active and reactive power, as shown in the following expressions:

$$P = \frac{V_s V_r}{Z_s} \cos(\delta + \theta) - \frac{V_r^2}{Z_s} \cos \theta \tag{2.6}$$

$$Q = \frac{V_s V_r}{Z_s} \sin(\delta + \theta) - \frac{V_r^2}{Z_s} \sin \theta + V_r^2 \frac{\omega C^{\text{line}}}{2} \tag{2.7}$$

However, the reactive power components are small. Usually, they can be neglected without making a significant error for medium voltage cables and not long overhead lines (shorter than 100 km). Therefore, C^{line} can be neglected for simplicity and equation (2.7) can be written as follows:

$$Q = \frac{V_s V_r}{Z_s} \sin(\delta + \theta) - \frac{V_r^2}{Z_s} \sin \theta \tag{2.8}$$

Both obtained expressions are equivalent to the transferred power expressions presented in classic transmission power systems, in which it is assumed that the X/R ratio is high ($X >> R$) and $\theta = 90^\circ$. This is presented later in this section and compared with other possibilities in distribution system, in which this assumption can change and the X/R ratio could have different values.

The equations (2.6) and (2.8) can be used for defining the voltage behaviour in terms of active and reactive powers. To simplify the analysis, it is assumed a unity impedance between both nodes ($\mid Z_s \angle \theta \mid = 1$) and the magnitude of the node s ($\mid V_s \angle \varphi_s \mid = 1$), both values in per unit. Therefore, equations (2.6) and (2.8) can be rewritten in term of per unit values as follows:

$$P = V_r \cos(\delta + \theta) - V_r^2 \cos\theta \tag{2.9}$$

$$Q = V_r \sin(\delta + \theta) - V_r^2 \sin\theta \tag{2.10}$$

Squaring and adding (2.9) and (2.10) the new expression corresponds as follows:

$$\begin{aligned} P^2 + 2PV_r^2 \cos\theta + V_r^4 \cos^2\theta + Q^2 + 2QV_r^2 \sin\theta + V_r^4 \sin^2\theta \\ = V_r^2 \cos^2(\delta + \theta) + V_r^2 \sin^2(\delta + \theta) \end{aligned} \tag{2.11}$$

which can be rewritten as

$$P^2 + Q^2 + {V_r}^2 (2P\cos\theta + 2Q\sin\theta) + {V_r}^4 = {V_r}^2 \tag{2.12}$$

$${V_r}^4 + {V_r}^2 (2P\cos\theta + 2Q\sin\theta - 1) + \left(P^2 + Q^2\right) = 0 \tag{2.13}$$

As shown in (2.13), the relationship between voltage and power is complex and non-linear. That brings two possible solutions for the voltage, as shown in the following expression:

$$V_r = \sqrt{\frac{1}{2} - (P\cos\theta + Q\sin\theta) \pm \sqrt{\frac{(2P\cos\theta + 2Q\sin\theta - 1)^2}{4} - (P^2 + Q^2)}} \tag{2.14}$$

In a similar way as presented in the literature previously shown, it is presented the voltage collapse as the main voltage instability problem in distribution system, where there is an operational point in which voltage starts to decrease uncontrollably. Solving the equation (2.14) or finding the conditions in which voltage instability is achieved makes the analysis more straightforward, by using any of the traditional methods (P-V curve, P-Q curve, or voltage stability indices)[43, 47–56]. However, this only gives some relevant features associated with loadability of each node, which normally contributes on surpassing the lower voltage limit defined in many operational standards.

Nowadays, voltage stability is a major concern in the planning and operation of modern distribution systems. There are considered the behaviour of Distributed Generation Units (DGs) and the interactions of both continuous and discrete protections and controls. A voltage instability in a distribution system is the possible tripping of its DG by its protection systems and the loss of load in an area [48]. Therefore, voltage stability analysis cannot be focused only in one parameter

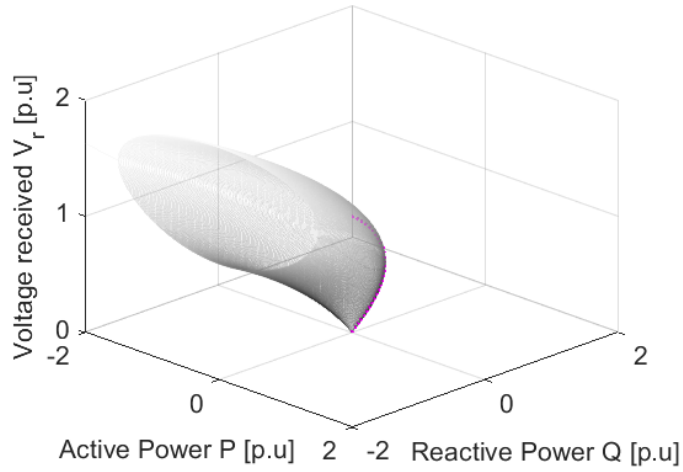
[57].

Analysing equation (2.14) solution is troublesome if changes in all variables are done simultaneously. Consequently, different scenarios are required in order to understand the possible situation that can be presented. The most basic analysis can be done with the transmission system's conditions, in which $\theta = 90^{\circ}$. The equation that represents that particular scenario corresponds to the following:

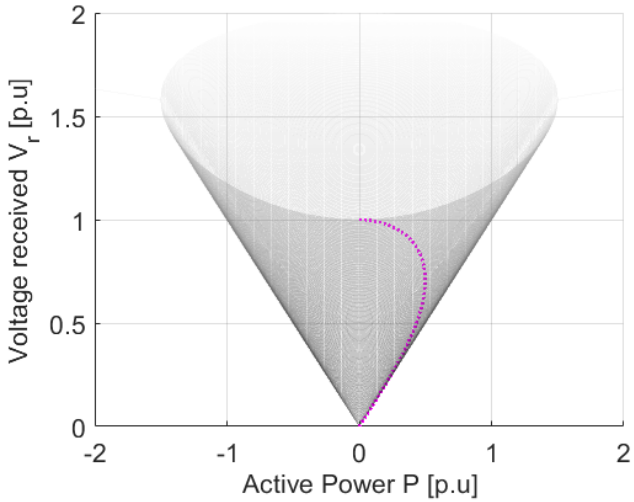
$$V_r = \sqrt{\frac{1}{2} - Q \pm \sqrt{\frac{1}{4} - (Q + P^2)}} \tag{2.15}$$

According to equation (2.15), there are two possible solutions after replacing the values P and Q. The one with the positive sign is called high voltage solution, and the other with the negative sign is called low voltage solution. Thus, to understand voltage instability according to this expression, load perturbation is assumed ($P \rightarrow P + dP$ and $Q \rightarrow Q + dQ$). On the one hand, a decrease in demand leads to an increase in V_r for the high-voltage solution. For in the limit that demand falls to zero, V_r tends to V_s and no power is transferred across the line. On the other hand, a decrease in demand leads to a decrease in V_r for the low-voltage solution. For in the limit that demand falls to zero, V_r tends to be zero. In general, there is a voltage-dependent component of the load (i.e. impedance), and therefore what will happen in practice is a divergence away from the equilibrium receiving-end voltage (i.e., $dQ < 0$ implies V_r falls which causes Q to fall further, and so on.).

An instability problem is observed in the low voltage solution. A drastic change in the system dynamics will be presented when this point is reached, which in the literature is presented normally as voltage collapse [58–64]. To illustrate this situation for the quasi-dynamics contexts, Figure 2.3 shows possible solutions of the equation (2.15) over the "cone" surface. During the sample of measurement, it can be assumed a constant value of power, but this value can change for the next sample (in this case, it can be assumed that measurements are accessible in more than a minute). Therefore, power and voltage are moving over this surface during different sample periods. In this figure, it is shown that the possible solution of the system equation where $Q = 0$ produces the magenta dashed line that can be projected in a separate plane. This helps to constructs the family of P-V curves between nodes, which corresponds to the classic curves for different values of P. The voltage instability occurs when the operation point moves close to the "nose" of this magenta dashed line, in direction to the low-voltage solutions presented before. In an analogous way, the same analysis is applied to produce the VQ curve of the node s with the projected magenta dashed



(a) Overview of the voltage solution surface for two-nodes system

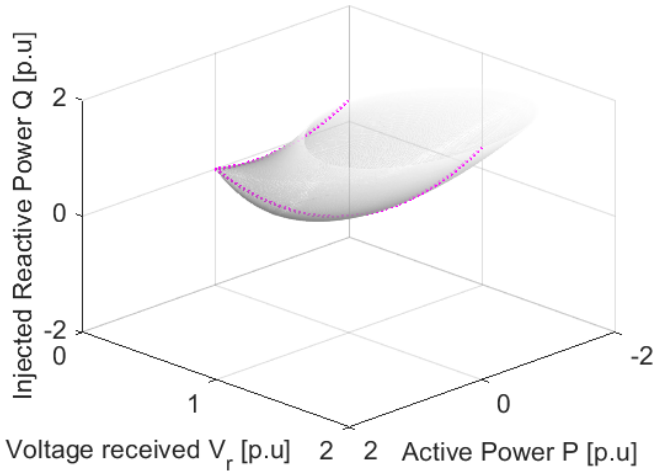


(b) View of voltage solution surface from the plane voltage-active power, which shows the classic P-V curve as a projected solution

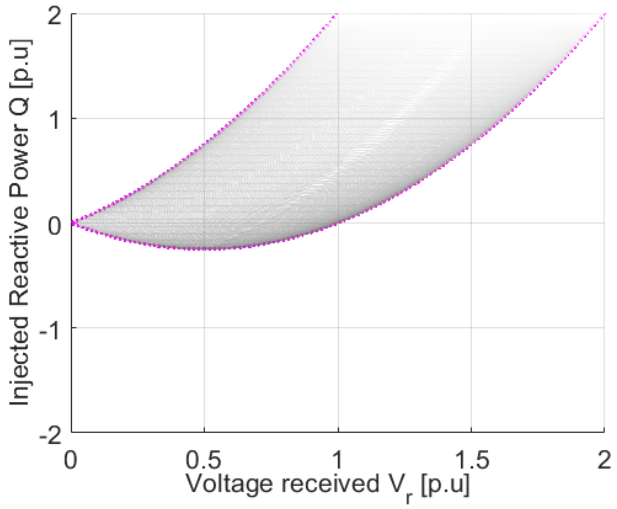
Figure 2.3: Voltage solution surface of equation (2.14) and the P-V curve projected from surface when $Q = 0$ and $\theta = 90°$

line to represent the solutions of the voltage equation, in this case, for $P = 0$. This is shown in Figure 2.4. Similar analysis can be done when the power is transferred in the opposite direction due to the energy surplus.

In practice, most of power systems operate at the higher-voltage solutions of equation (2.14), so have some inherent stability. Then, the voltage has to drop a lot before "flips" to the low-voltage solution, and operational limits usually rule this out. This thesis focuses on the phenomenon around the operational point when



(a) Voltage solution surface considering the injection of reactive power



(b) View of voltage solution surface from the plane injected reactive power - voltage, which shows the classic VQ curve as a projected solution

Figure 2.4: Voltage solution surface and the VQ curve projected from surface when $P = 0$ and $\theta = 90°$

uncertainties are considered due to the balance generation-consumption. The consideration of disturbance due to the nature of renewables and load fluctuations makes that the voltage oscillates around the operational point [65]. Changes in the operational point will not be big enough to put the system in an unstable point (that means, the basin of attraction from expected operational points can absorb a perturbation without shifting to an alternative state). The case under study does not represent an abrupt transition shifts in the voltage when chan-

ging conditions pass a bifurcation point, since the voltage is oscillating around the operational point and the allowed operational limits. Therefore, the voltage will always reach the nominal value when the generation (due to renewable units) and loads get balanced in the analysed node, as indicated in Figure 2.5. That helps to conclude that the voltage problem analysis in the distribution system does not respond to a voltage stability problem considering this quasi-dynamics in the traditional way if the desired modelling and control will be in a regime close to the desired operational points. This simplifies the analysis and drives the research to develop a model (ideally linear), that represents this relationship and reduce the complexity of the equivalent model seen in equation (2.14).

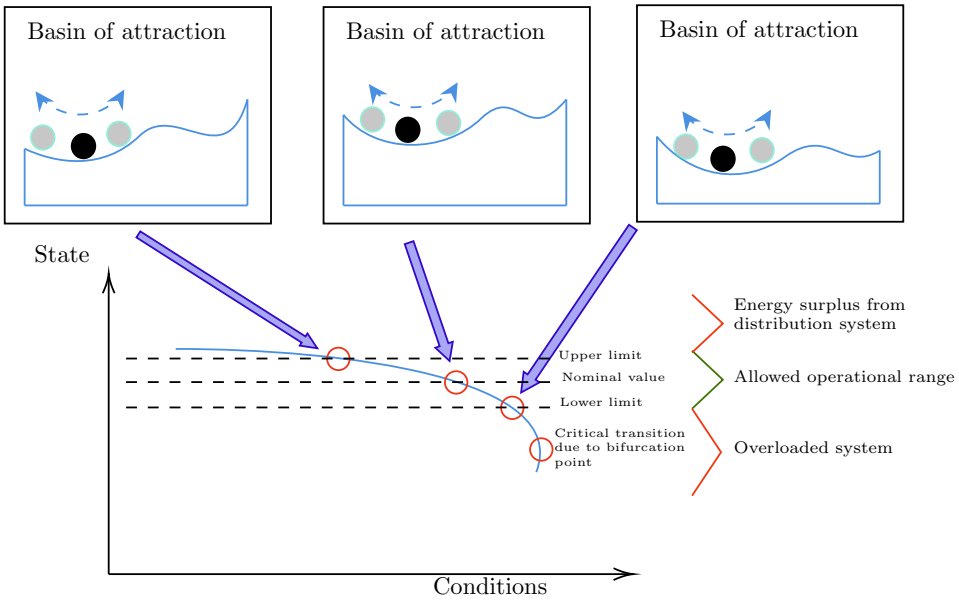


Figure 2.5: Illustration for voltage changes due to increase or reduction

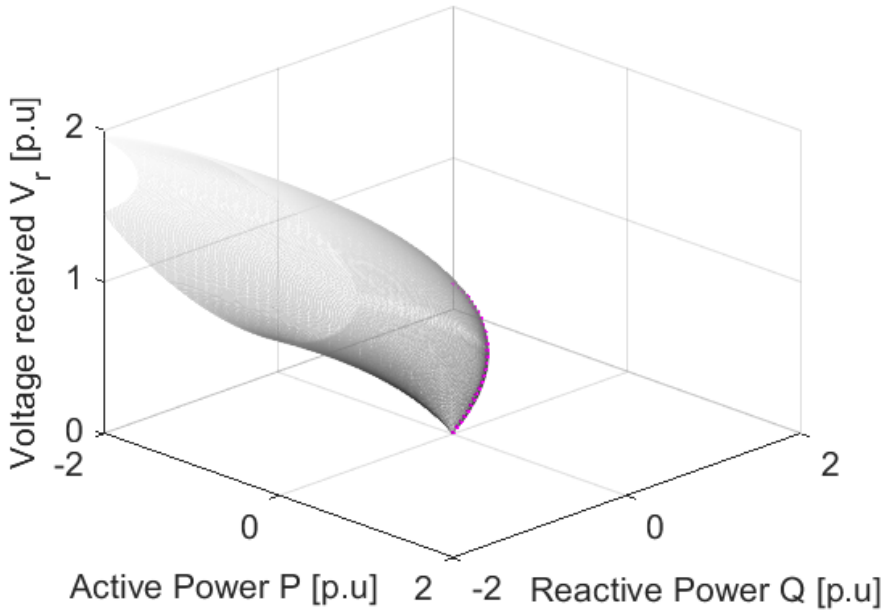
It is assumed the X/R ratio is small for several distribution systems and actuators are focused only on active power due to the higher resistive component R in most of the grid. However, this assumption can be quite general and discard the analysis of each active and reactive power's impact. To illustrate this, Figures 2.6 and 2.7 show an example of the effect of changing the value of θ with higher R that corresponds to an angle impedance in the particular case of 75°. For a fixed X/R ratio seen from the node V_s, there are two possible scenarios for the active power action: the total balance of active power is negative, where the node is consuming power (which is reflected in the impedance seen from one node to the other and produces that $\theta = 75°$), and the opposite when the balance is positive, where the node is injecting to the system (the angle seen from one node to the other is then $\theta = 105°$). Figure 2.6 shows how the surface in both curves is switched from

one plane to the other, which means that the range of action and the impact of P increase when the observed θ increases. That is expected when the system has a high resistive component R, and the node provides power to the system. Figure 2.7 shows how for both scenarios, when active power is consuming or injecting, the margin of action for the reactive power Q remains within the same range. A similar analysis can be done as was presented for active power to understand the node's behaviour when reactive power is consumed or injected. Therefore, even for impedance between nodes that are different from the general assumption of X/R ratio, the shape of surface solution remains the same. In fact, it is important to highlight how possible solutions keep the same shape while is swift when X/R ratio and the node is consuming/generating. It is particularly important to analyse this problem in distribution system with a different assumption from the traditional value of X/R ratio, especially if a linear approximation is desired to model the relationship between power and voltage.

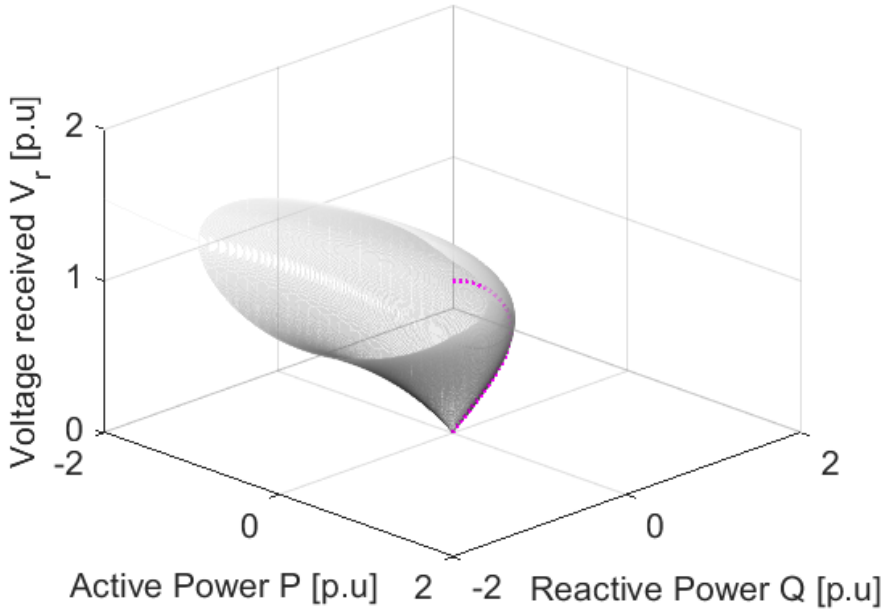
This analysis introduces the idea of modelling the distribution system beyond general assumptions that does not fit with the current status of the network. It is concluded that the voltage modelling (and potentially controlling) can be approached from a different perspective, considering that the distribution system is operating close to a stable operational point from the classic voltage stability perspective. Therefore, a new challenge would be to obtain a good representation of relationship between voltages and active and reactive powers, that would be constantly shifting between generation and consumption and capture the transition in between. Also, this representation of distribution system must help to maximise the efficiency of system operation and consider the fluctuations of the operational conditions for distribution systems. This must be done beyond of assuming general conditions in the network such as fixed low X/R ratio, that does not apply anymore or are not able to represent correctly the nature of these distribution system quasi-dynamics. This idea is the base for the preliminary approach presented in section 4.4 and particularly in section 4.4.2.

2.3 Voltage control for distribution systems with high penetration of renewables

In the previous section, it was introduced the regime of possible voltage solutions for different operational conditions and contrasted them with different assumptions, such as the X/R ratio that differs from the classic ones (reflected in the variation of θ). The main goal is to establish the fundamentals required to understand what the voltage operational goal would be to achieve once the distribution

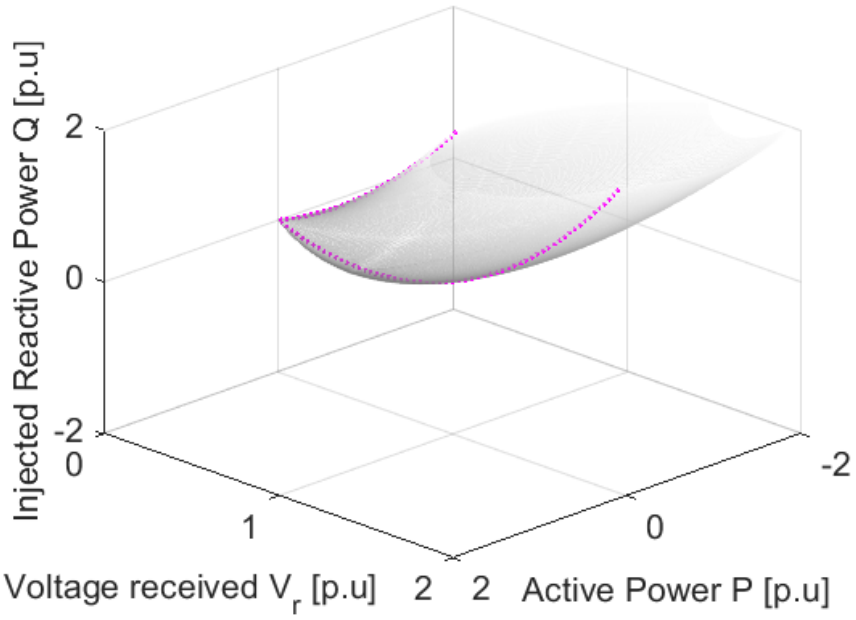


(a) Voltage solution surface of equation (2.14) when $\theta = 75°$ (X/R ratio represents the angle of 75° and active power is consumed in the node)

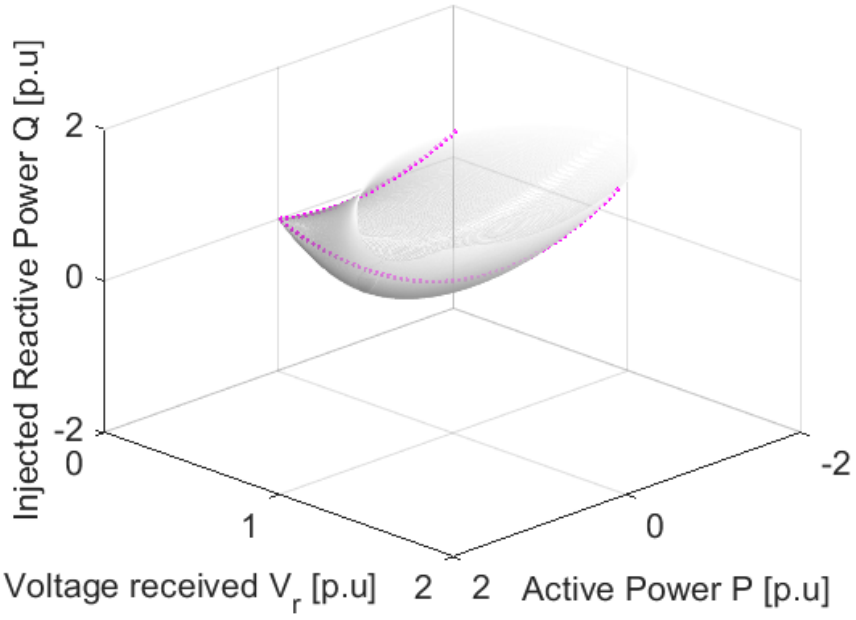


(b) Voltage solution surface of equation (2.14) when $\theta = 105°$ (X/R ratio represents the angle of 75° and active power is generated in the node)

Figure 2.6: Voltage solution displacement when θ is increased and effect on the P-V curves



(a) Voltage solution surface of equation (2.14) when $\theta = 75°$ (X/R ratio represents the angle of 75° and active power is consumed in the node)



(b) Voltage solution surface of equation (2.14) when $\theta = 105°$ (X/R ratio represents the angle of 75° and active power is generated in the node)

Figure 2.7: Voltage solution displacement when θ is increased and effect on the VQ curves

system representation is obtained and to guarantee good operation in the network. So far, it has been discussed the similarities and differences between understanding voltage in distribution systems and the classic voltage stability problem, as well as the surface solutions obtained for scenarios with typical assumptions in transmission systems and how these can change with different assumptions in distribution systems. In modern distribution systems with high integration of renewable energy sources, the problem is slightly different compared to former radial distribution systems [66, 67], as this assumption can be constantly changing. During some periods of the day, there may be an energy surplus that flows to the rest of the network when generation surpasses energy demand at the connection point. Distribution systems were not initially designed to provide energy to the rest of the system, which can cause over-voltages and damage different devices. Therefore, it is crucial to have a deep understanding of how these conditions are constantly changing from the model to help define a potential control approach.

Assuming that all measurable and available data are processed, and once that is constructed a model able to represent relevant internal states from the distribution system, it is possible to define the control target to maintain the voltage under operational limits. Further restrictions, such as component loading and power balance constraints, may be imposed in addition to the voltage variations considered in the proposed approach. Nonetheless, the methodology employed for modelling the distribution system and identifying potential control strategies can be extended to incorporate these additional constraints.

The control stages that applies for this problem can be summarised in data acquisition, system identification (construction of plant and definition of constraints), system analysis and control actions. The first two steps are related one to each other. An identification of the system quasi-dynamics is required, considering the partial observability in distribution systems. In this scheme, uncertainties are presented as disturbance that must be considered in the robustness scheme of the control strategy. Several control challenges for this problem include the coordination and control of heterogeneous components (OLTCs, capacitor banks, and Static VAR Compensators (SVCs)) that possess different timescales, so as to lead to enhanced operation with minimal participation from the operators during normal operation. Also, it is expected an efficient control of reactive power and voltage control using distribution-level power markets by providing incentives for flexible loads and distributed resources and transient stability, frequency, and voltage control in the presence of islanding in microgrids, among others.

To tackle this variability, the most widespread practice to achieve acceptable operational conditions consists of curtailment based on local measurement [68].

This results in a practical solution in which inverters stop injecting power into the system. However, this reduces the margin of renewable energy sources that can be integrated into the distribution system. For some countries, this can also become a source of conflict between customers, since compensation schemes based on the energy provided to the system can be unfairly defined according to operational conditions between customers who are close to each other [69]. Therefore, coordinated voltage control that improves the use of renewable energy sources in the distribution system is required. Nowadays, elements such as capacitor banks, batteries, and regulators can be used under the right control scheme to improve this integration.

A possible recommendation for the suggested approach could be the integration of uncertainties as inputs instead of disturbances. This is expected to avoid procedures such as curtailing the energy injection to the system, which means increasing the flexibility of the control strategy to deal with possible scenarios instead of assuming worst cases as is usually done. What does it mean to consider something as an input rather than a disturbance? It suggests something that can be directly manipulated, especially using feedback, to obtain desired outputs. A disturbance normally represents an exogenous signal that cannot be controlled.

Only with the purpose of illustrating this idea, the following quasi-dynamic system presented in this section represents the distribution system with all its exogenous variables, and it is represented by matrices A, B and E in the equation

$$\frac{dx}{dt} = Ax(t) + Bu(t) + Ew(t) \tag{2.16}$$

where $u(t)$ is an input and it is assumed that it can be changed according to the state $x(t)$ or $w(t)$ ($u(t) = g(x(t), w(t))$), whereas $w(t)$ is just some uncontrollable signal that takes values in a given set.

Unfortunately, curtailment seems to fall under the bracket of regarding the source as an input, because it manipulates the power injection based on current needs (i.e. the state). A key distinction is that a more systematic and effective approach is required, rather than relying solely on curtailment. This involves integrating control as a feedback law like $u(t) = g(x(t))$ considering, for example, the quasi-dynamics

$$\frac{dx}{dt} = f(x(t), u(t), w(t)) \tag{2.17}$$

rather than making decisions on curtailment levels based on static power flows. An alternative view of curtailment could be to regard it still as an exogenous input but one that is partly controlled by just attenuating its contribution. For example,

the quasi-dynamic system can be represented as follows:

$$\frac{dx}{dt} = Ax(t) + Bu(t) + E(x(t))w(t) \tag{2.18}$$

where $E(x(t))$ (the gain between $w(t)$ and the system) is varied by curtailment strategy. Here the input $w(t)$ itself is not changed, but its amplitude is. More generally, $E(x(t))$ could be regarded as a filter.

One of the research questions after reviewing different papers concerns understanding voltage control for distribution systems with high integration of renewables, which implies an understanding of the relevant dynamics from this part of the system. The literature review shows two main philosophies: the first one is more focused on the power system assets' capabilities and the measurement available in the distribution system. This approach has been used for years because it is a practical and feasible solution. Relevant examples of this control branch are presented in [5, 69–73]. Normally, this control is simplistic and rule-based for distribution systems, in which the regulation defines the voltage margin in the distribution system according to the voltage level. However, there is no dynamics analysis behind this approach. This rule-based heuristic approach has several disadvantages, including the limitation in the amount of power that can be injected into the distribution system from renewable energy sources, limiting the voltage action in non-scalable schemes, and short-term and long-term voltage collapses, among others [74]. The other branch of voltage control is more theoretical and understands the voltage stability from control theory, in which the voltage dynamics are written down and advanced analysis/design of the system is done. The level of control theory content is high, which differentiates this family of papers from the works of the previous branch. Relevant examples are shown in [40, 54, 56, 75]. However, the papers are exclusively about microgrids and inverter-based control or transmission networks and reactive power control/optimisation (environments in which most of the internal states are observable). Therefore, this review highlights a gap that represents a good opportunity for this thesis to apply control theory in the distribution system, which considers partial observability and uses the capability of installed assets to maintain voltage within operational limits.

A natural strategy for controlling in this kind of environment is Model Predictive Control (MPC) approaches because operational limits are key in a power system. This technique predicts the system quasi-dynamic considering constraints, which makes this approach an interesting candidate to be applied. There are some experiences for voltage control in power system applications. In [76], the control scheme was proposed to manage the load shedding and maintain the voltage in the required range for a transmission system. A quasi-dynamic model for loads

and the voltage constraints were integrated into the control algorithm. In [77, 78], fast MPC is applied to control a specific voltage by controlling an inverter associated with a substation. In [79], MPC is used to define a Voltage/Var Optimization (VVO) scheme and control capacitor banks and OLTC for a distribution system with DGs considering uncertainties associated with renewable energies. Similar work has been done in [80], in which the main goal was the optimisation of voltage variation reduction by only considering OLTC. Recent works have shown the integration of other devices such as Energy Storage Systems (ESSs) [81, 82], and different time-scale control modes have been defined to develop the voltage control based on sensitivity coefficients [83]. Even there are works developed for centralised [84] and decentralised MPC [85, 86], which can improve considerably the impact in the overall control scheme. MPC offers a good compromise between rigorous control theoretic analysis and synthesis and practical design. Therefore, it fits the specific requirement and could bridge the gap identified in this thesis, where on the one hand you have practical rule-based approaches with no guarantees and on the other hand complex control law designs based on first-principles models.

One disadvantage of using MPC controllers is their high dependency on accurate system models for effective control. This can be problematic for several utilities in distribution systems where the accuracy of models obtained from non-updated databases may be limited. Data-driven approaches offer a solution to this problem by complementing the MPC approach and capturing quasi-dynamics within the system data, such as voltage, current, and power, as demonstrated in [87] which used Auto-Regressive-Moving-Average Model (ARMA) for controlling power electronic converters connected to the electric power system. While model approximation may not adversely affect the quality of MPC regulation, it is essential to check for voltage stability problems in distribution systems.

Finally, the MPC approach has shown a significant impact in this type of problem because of its nature as an adaptive scheme considering constraints. Even though there are not many applications focused on voltage control for distribution systems, an interesting research challenge would be to combine system identification approaches to obtain a more accurate model and to analyse the uncertainties that probably affect the robustness of the control. For this thesis, the MPC approach was not implemented since it is not part of the covered scope. However, it is important to highlight the fact that the nature of the controller design based on models suits perfectly with a data-driven approach, which will give a reference of how the proposed modelling should be developed: a model that captures what is happening in the distribution system based on measurable information and not on

any estimated/approximated parameter. Therefore, it is expected that any MPC approach integrated for controlling voltage, in this case, fits perfectly with a data-driven model. It could be an interesting research question once a good data-driven model is obtained, which is in fact the actual focus of this thesis.

2.4 Observability and controllability in distribution systems

To achieve the control goal presented in the previous section, it is necessary to explore how well the internal states of the distribution system can be represented based on what can be sensed or measured from the system. Therefore, these states must be inferred from the measurements of some variables, as well as evaluate how easily system states can be changed with insignificant risk and effectively. This corresponds to the observability and controllability problem in the distribution system. The former represents the capacity to measure the current internal states of the system by only using information from outputs (e.g., voltages in some nodes, status of tap changers and circuit breakers, and some exogenous information such as census or weather), while the latter is the capability to transfer these internal states to any particular state, in a finite time duration, when a controlled input is provided to it (which is mainly associated with power injections or consumption, e.g., load consumption, renewable energy) [88].

The assumption of full controllability and observability in the transmission system is commonly made because the available data, models and actuators can sense and manipulate inputs in the system that excite all modes of interest [89, 90]. Unfortunately, in the distribution system, this is not the reality due to the excessive costs associated with adding sensors and actuators over the whole network. Therefore, there is limited knowledge of the system topology and parameters and limited availability of measurements. Relevant questions arise from this, such as *which* nodal voltages are relevant to represent the internal states that represent the voltage quasi-dynamics, or if it would be possible to use measurements or any *metrics* that help to represent these quasi-dynamics. Additionally, once the distribution system is represented, it is necessary to evaluate the impact of power injection/consumption to excite the internal states and bring them to a desired level, even if not all of them are accessible through measurement.

Recent research on distribution systems, in order to represent their internal states through classical electric modelling in the absence of complete information, has focused on the state estimation problem only [91–98]. These approaches rely on a sufficiently accurate system model that maps states to measured data. In

several real distribution systems, where the number of nodes and lines may be large, this model and its parameters may not be known. This has motivated the use of grey-box modelling approaches to estimate the unknown parameters in the physical model [99, 100], but even so, a common assumption is the knowledge of the system topology. A further limitation common to state estimation approaches is the technical requirement of system observability. In practice, this leads to the requirement that there should be sufficient measurement units around the distribution system to allow accurate state and/or parameter estimation. Unfortunately, in many distribution systems, measurement units are not prevalent or widespread [101]. Accepting this reality, an alternative question and perspective contemplated in this thesis is: considering an almost complete lack of knowledge of the system topology and parameters, what kinds of models could be created by considering the measurements of system variables that are available? Furthermore, what information can be extracted regarding the controllability and observability of the distribution network? Which nodes are most useful for observation and control in the context of managing the system voltage?

A possible solution to overcome these issues is to use a data-driven approach, which can help to develop models in distribution systems with limited observability, without relying solely on classical, non-scalable physical-based models [93, 96]. The next section explains in more detail how data can be used to address the issue of observability in distribution systems.

2.5 Data-driven modelling of time-series background

Previous section introduced the observability and controllability problem in distribution systems, and also the idea of using data to overcome and improve this condition. Several data-driven approaches to power system modelling and estimation assume that the system is fully observable, which in turn implies that the measured data reflects the time evolution of the modes to be analysed [88, 102]. Correspondingly, in many power system control approaches the assumption of full controllability is made, which in turn implies the manipulable inputs in the system are able excite all modes of interest [89, 90, 103, 104]. Therefore, an alternative problem to be considered is the scenario where there are limited knowledge of the system topology and parameters, and limited availability of measurements [105].

In this section, it is explored which data can be used and the potential to improve the partial observability and controllability condition in distribution systems. The main challenge is that it is impossible to measure everything in distribution systems, yet it is desirable to have an indication of the current state of

the network and its proximity to a critical state. This has traditionally been done in the context of voltage stability, and several effective and practical approaches have been developed for calculating a "voltage stability index" from available data. However, a new metric is required now that the problem is different from voltage stability.

2.5.1 Requirement of new metrics to develop model

As mentioned earlier, it is required to explore whether there is any available data that can provide insight into the voltage quasi-dynamics, without accessing measurements of all internal states of the distribution system model. Similar to this approach, the voltage collapse problem can be understood through measurable parameters that provide an idea of how strong the system is to deal with this particular condition. To tackle possible voltage collapse problems, the use of voltage stability indices has been proposed by several authors [47, 52–56]. In [56], a voltage stability index for distribution systems is presented, which showed satisfactory performance for traditional radial distribution systems. This index serves as a good indicator of voltage collapse as it is represented by a scalar value. However, the performance of this index under new operational conditions, including DGs, was not discussed or well-supported.

In [52], a dynamic analysis for voltage stability was presented considering the real behaviour of the system, such as loads (dynamic and static models), DG units, automatic voltage and frequency control equipment, and the protection systems. The overall power system was represented by first-order differential equations. The static analysis was conducted by analysing small-signal stability, which is achieved in the frequency domain using eigenvalue analysis. In [55], the authors propose obtaining voltage sensitivity analysis for both active and reactive injection in the distribution system and modelling the Distributed Energy Resource (DER) units with a voltage regulation model to deal with the uncertainty of load/generation over the distribution system for voltage control. The results showed beneficial under the given voltage regulation target and give full play to the renewable energy power plant reactive power control capacity.

In [53], an easy-to-implement voltage stability index is proposed, based on load characteristics, distribution system specifications, and DG characteristics. The index contains critical data on the distribution system node voltages, such as voltage stability status and sensitivity of the voltage to power changes. This index can rank the system nodes based on their voltage profile and voltage stability status, and it can be used as a practical tool to identify the best candidates for installing new DG units. The goal is to improve voltage stability and elevate undervoltages without

causing overvoltages in any of the system nodes.

Since traditional metrics that focus solely on the voltage stability problem rely on prior knowledge of the system, particularly the topology configuration, additional measures are needed to assess the impact of DG units, taking into account the natural behaviour of the system, such as load consumption profiles. The new metrics must be capable of capturing information about the system, even under conditions that are far from instability. They should not only reflect the static overall condition of the system but also adapt to the operational state of the system. Therefore, new metrics that concentrate on operational conditions far from instability need to be explored, describing the system with the available data.

The first natural metric used for this purpose would be the power injected into each node, since it can provide some insight into how much the voltage may change according to its changes during different periods of time. Therefore, predicting how these injections/consumption patterns change over time could give an idea about voltage variability. However, this occurs in different locations around the distribution systems, and it may not improve or have an effect on a voltage variation at a specific point. Therefore, exploring the impact of power across the system, without relying on system topology or proximity, can be considered. A similar approach to this idea is proposed in [106], where some power values were enhanced using the Fisher-Z transform and relying on some topological parameters for frequency control applications. This kind of approach has not been applied to modelling, which could be a good reference to investigate and explore the application in modelling approaches.

There is another metric that is not commonly associated with these modelling approaches. For instance, the distance between measurements is a spatial-temporal characteristic that gives a notion of 'electrical distance' between nodes. This can give a sense of how far two nodes are located and provide a measure of to what extent a voltage variation at one node can be inferred from variations at another node. In that sense, it can determine whether the nodes are similar or not in terms of observed voltage variation and, therefore, whether voltage needs to be monitored or controlled at just one or both nodes. The literature presents different works analysing the correlation in voltage to describe this metric and propose a topology of the system when it is unknown [107–109]. The challenge, however, is to estimate the electrical distances between different nodes in the network from measured data.

These are metrics that are able to explain aspects of the system and its assets, such as the location and size of perturbations. They should be measurable and stored in time-series data, and potentially used as inputs for a system identifica-

tion approach, not necessarily as control inputs, but as parameters that describe what is happening in the system. This represents a novel approach that could be achieved by using these metrics in a modelling (and potentially controlling) framework. Therefore, it is required to develop a model (preferably linear) that incorporates these new metrics that use the available data to "construct" a picture of the distribution system. This representation should not rely on iterative calculations or conventional power flow solutions, which require substantial prior knowledge of the electric system. As a result, it can be used in real-time applications, which is a significant advantage for distribution systems with thousands of nodes.

2.5.2 Distribution system modelling and system identification

2.5.2.1 Static vs Dynamic models

Once the metrics are determined, the next step for the voltage analysis corresponds to the modelling of the distribution system. According to the period that this analysis is conducted, modelling parameters and analysis procedures are executed for steady-state events [53, 110] and dynamic events [111–113].

Static modelling and analysis are intended to evaluate the system after it has transitioned to a new operating state, and to avoid the study of transients in between the two operating states. Therefore, these analyses are done based on traditional power flow methods, including the construction of the admittance matrix. Dynamic modelling and stability analysis are intended to evaluate the transient voltage behaviour of the distribution system in between two steady-state operating modes. Therefore, additional characteristics associated with device controllers must be considered.

The latter is more complex, and there are few methods presented in the literature to obtain the Electromagnetic Transients Program (EMTP) solution [53, 114–119]. In [53], both static and dynamic voltage stability analyses of distribution systems are performed using a voltage-dependent load model. Additionally, a method based on finding an EMTP solution in the shifted frequency domain and then transitioning the solution back into the real-time domain is presented [119]. The method, called Shifted Frequency Analysis (SFA), reduces computational calculations effort by finding an equivalent DC system that represents the dynamics. To clarify and simplify the problem under study, this thesis considers only loads modelled as constant power, and the analysis of the dynamics under study will be based on different sources, as explained at the end of Section 2.5.4.

For system planning and operation, dynamic analysis becomes particularly

important when devices such as generators, inverters or SVCs control participate in the system. Additionally, the variability of some renewable energy sources such as PV units may cause unpredictable operational changes that might result in dynamic voltage instability problems due to the dynamic behaviour of inverters and step-up transformers [96].

Therefore, voltage stability analysis must include both static and dynamic analysis to guarantee a better understanding of voltage control in the distribution system under continuous changes of operational conditions.

2.5.2.2 Physics-based models vs input/output models

For both static and dynamic models explained before, there is a shared requirement of a complete knowledge of the system's parameters for the modelling and calculation approach. Typically, power system modelling in the form of Physics Based Models (PBMs) is used, which involves using devices' characteristics that have a physical meaning or can be obtained directly from the observation associated with the analysed phenomenon. In power systems, this method is applied, for example, to model the connections between different nodes, loads and generator units through lines, which correspond to the resistance and the inductance of the real line that makes this connection. Unfortunately, several utilities have not updated the information of all their assets, and some parameters are difficult to obtain due to the size of the distribution network. Therefore, a data-driven approach that considers updated measured data and defines the control target is required to improve the observability of the distribution system.

[120] is an old paper that predates the smart grid paradigm by some time. In that sense, it was visionary in anticipating the need to model the network from data. The approach proposed in this paper involved injecting an impulse signal to excite some dynamics of the system and construct a model based on the response obtained. Similarly, in [121], a harmonic generating device was used to inject a current of 100A peak, and the voltage was measured for different frequency values. Both methods accurately identified the distribution system with relatively simple hardware. However, special equipment for a disconnected distribution system was required to apply this method. A disconnected system cannot provide a good understanding of the system dynamics, especially when renewable energy sources are connected. These works highlight the importance of understanding the different analysis methods when identifying a system that can evolve in space and time.

Recent research related to distribution system data is mainly focused on distribution system state estimation [91–95, 97, 99, 122, 123], from which the steady-state

and dynamic states of the distribution system are obtained. With this information, it is possible to reconstruct certain data from the system. For instance, the use of real data obtained from Intelligent Electronic Devices (IEDs), such as protection relays or Phasor Measurement Units (PMUs) is shown in [99]. In [91–93], a real-time system model is obtained using information available from PMUs based on the inverse power flow problem. The full admittance matrix can be obtained when the system is completely observable (without hidden nodes). Nevertheless, a partial admittance matrix can be obtained from a partially observable system (with hidden nodes) by using Kron reduction and complementing with a proposed algorithm based on graph theory. In [94], a review on distribution system state estimation approaches based on data is presented. There are several methods, most of them based on Kalman Filters, such as Weighted Least Square and Iterated Kalman Filter (IKF) methods integrating PMUs[123], Extended Kalman Filter (EKF)-based State Estimation, Unscented Kalman Filter (UKF) method [122], State Estimation based on Ensemble Kalman Filtering [95] or Koopman Kalman Filter [97]. The main disadvantage of these approaches based on state estimation is the observability assumption of the system. The system model is more accurate, and the variables are estimated as much as there are several measurement units around the distribution system, which helps to construct the real status of the system. Unfortunately, several utilities do not have enough measurement to estimate all the dynamics (and quasi-dynamics) of the system.

In [19] a discussion between PBM and Input/Output Model (IOM) was presented. On one hand the PBMs are easier to develop and understand in power system applications. However, PBMs have difficulties to get high order dynamics and the use of more than one model to get different dynamics increases the computational effort. Additionally, estimation of several parameters is not beneficial and PBMs are better when the system has only a type of components such as generators, loads, motors, etc. On the other hand, the IOMs describe external systems and they are focused on the input/output characteristics rather than its physical structure inside. This helps to get an updated model with the current characteristic of the system dynamics, including the load fluctuation and the variable power injected by renewable energies. For the latter, recent method like the Nonparametric System identification of Stochastic Switched Linear Systems presented in [124] can be applied, which tries to construct a good approximated low order model considering a noisy input-output model. This method uses a probabilistic approach to consider the switches as exogenous but not as control input, which is useful when voltage as continuous state is defined as a function of a factors like solar radiation.

A similar work to this thesis that considers data as the source to define voltage

set-points in the control strategy is presented by Cupelli et al. [125, 126], which use functional stochastic gradient descent in Reproducing Kernel Hilbert Spaces (RKHSs), a machine-learning-based method that expands the dimension of the model to obtain linear relationships. The main disadvantage of this approach is that the dynamics within the model can lose the meaning that explains how variables are related.

For some system identification approaches, it is assumed that the system is fully observable, which implies that the measured data reflects the time evolution of the modes to be analysed [19, 34, 127]. Additionally, the assumption of full controllability is made, which implies that the inputs can excite all modes of interest. This represents one of the most significant challenges of this thesis since measurements are not available in every single node of the distribution system, and the options for sending excitation signals in the distribution system are limited.

In [128], an online virtual metrology of distribution line impedance was used to calculate parameters from the voltage drop linear equivalent computing model. The equation to obtain the parameter model can be solved by regression analysis, average value of solving equations method or smart algorithm based on Artificial Neural Networks (ANNs). This model worked considering partial observability for a radial distribution system; however, its use was not stated for identifying a distribution system that includes renewables or any generation unit. Additionally, the linearity of the system must be guaranteed, and validation for a meshed system was not performed. Finally, an ANN-based approach needs to be installed in power system equipment, and the selection of the ANN topology does not follow any procedure.

A discussion of this is presented in [129] focused on the dynamic model representation of active distribution network cells and microgrids. Strong points and drawback of using conventional dynamic system reduction, Ward equivalencing, Modal Analysis and Coherency based methods are presented. Most of them avoid the integration of non-linearities, and the measurements-based approaches are presented as an alternative to deal with this. Using this information, black-box modelling approach is presented as an alternative by using ANNs (when there is no information of model parameters available) or grey-box modelling approach (when the available physical knowledge allows selecting a physically parameterised model structure). The latter approach considering grey-box suits perfectly for modern distribution systems, which are partially measured and partial information of some parameters such as resistance and inductance of lines are available.

2.5.2.3 Choosing a model structure for system identification

Choosing an appropriate model structure is a crucial part of the system identification approach [102]. There is no universally best system identification approach [103], which means that exploring different model structures is required. The Auto-Regressive Exogenous (ARX) model and state-space models can handle several model structures effectively because they have efficient algorithms [89, 90, 102, 104, 130].

In [19], the limitations of the ARX model for system identification in power systems were discussed, and it was argued that subspace state-space identification algorithms are more effective. The ARX model is highly dependent on the type and location of a disturbance used to excite the corresponding dynamic modes of the system. For distribution networks with DERs, such as Doubly-Fed Induction Generators (DFIGs) in wind generation, the dynamic equivalent model proposed in [19] did not show a significant advantage over the constant power model. However, the opposite was found to be true for distribution networks with several conventional synchronous generators. Therefore, the model proposed in [19] may not be suitable for distribution systems that contain diverse types of distributed generators, but this is an area that could be explored further, for instance, by integrating Eigensystem Realization Algorithm (ERA) and Observed/Kalman Filter Identification (OKID) [131–133].

In [134], various Auto-Regressive-Moving-Average Models with Exogenous Inputs Model (ARMAX) models were evaluated for estimating electromechanical oscillation damping, and all showed adequate accuracy with slight differences. ARMAX methods may have low-order model structures that reduce the computational burden, and the recursive calculation method was rapid during optimizing the model coefficients.

The previous polynomial regression structures can be converted into a state-space form, which gives more insight into the internal dynamics [102]. It is important to mention that this conversion process is not unique. According to Phan and Longman [135], the relationship between input-output data and the coefficients of an input-output model is linear. On the other hand, the relationship between input-output data and the state-space model parameters is non-linear. Therefore, there are several representations that can be obtained in state-space form that will capture the internal dynamics presented in the data used for modelling.

One alternative for producing state-space representations is based on using Koopman operator approaches [136, 137], which is a linear, infinite-dimensional operator that represents the action of a nonlinear dynamical system on the Hilbert space of measurement functions of the system states. This is useful for identify-

ing intrinsic coordinate systems and representing nonlinear dynamics in a linear framework. This procedure strongly relies on measurements, meaning that the Koopman operator does not depend on linearization of the dynamics. Instead, it transforms the measurement to an infinite-dimensional representation in the Hilbert space, and thus represents the dynamical system's flow on measurement functions as an infinite-dimensional operator.

In this approach, the challenge is to find the right Koopman operator that aids in this representation [31, 138]. Therefore, several procedures are applied to obtain an approximate finite-dimensional approximation of these operators. These system identification techniques produce low-rank state-space models using data-driven techniques such as Dynamic Mode Decomposition (DMD) [138–140], or extensions such as Extended DMD (eDMD) [141, 142] and Sparse Identification of Nonlinear Dynamics (SINDy) [143]. In addition to the previous methods, some nonlinear techniques can be applied to reproduce internal dynamics, such as Nonlinear Auto-Regressive-Moving-Average Model with Exogenous Inputs Model (NARMAX)[144].

Some applications have been attempted for power system models, in small portions of systems or power electronics models in stability analysis [145–148]. An interesting challenge is the application of these techniques to identify large systems such as distribution systems.

The main idea of this thesis is to find a procedure that can capture the most important dynamics around the operational point for voltage analysis based on information taken from the system. It is not expected to estimate the values of certain variables such as voltages, but to identify the actual behaviour and make conclusions regarding its actual operation that drives the voltage control target. Different model structures used in system identification must be assessed for this kind of problem, to obtain a low-order dynamic model that can be used for voltage control. A more precise discussion of the meaning of these "dynamics" (which differ from the traditional dynamics in power systems analysis) is extended in section 2.5.4.

2.5.3 Spatio-temporal system identification

In addition to the time characteristics discussed in the previous section, it is also important to consider the location of perturbations in the modelling approach [144]. The relevant location of the perturbation and the system component that may be impacted can be determined based on the existing knowledge of the system's topology and the rated values of load/generation units. However, in cases where there is limited information of the distribution system, the topology

of the system may not be directly inferred, or it may be unknown altogether. In such cases, the location of the system perturbation can provide insight into the selection of a specific node for voltage prediction or potential control.

According to Zhou and Buongiorno [149] and Martin and Oeppen [150], accurately incorporating the structure of spatial dependence into the model is a critical challenge. It is proposed splitting the problem into two primary tools for operationalising spatial dependence: spatial weight matrices and spatial lag operators, as well as the concept of the order of spatial neighbours. The order of spatial neighbours refers to their distance from a specific location. First-order neighbours are the closest characteristics to the location, while second-order neighbours are farther away than first-order neighbours but closer than third-order neighbours, and so on. For a regular grid system, a standard definition of spatial order is available; however, for irregular systems, the model builder must define the order of spatial neighbours [151]. Similar works consider the impact of these weights and articulate them with ARX or ARMAX model approaches, creating Space-Time ARX and ARMAX models [152, 153].

In power systems, several studies have employed the concept of location to construct prediction models [154–156]. However, these studies have used location as a clustering tool to focus on measuring the impact, instead of deducing location based on measured data. In Horak et al. [157], the correlation of location was compared with other features to evaluate the impact of model prediction based on spatial knowledge, while [156, 158, 159] proposed different indices or features based on space/distance to provide information about location and impact on the generated model. In this thesis, to reduce the complexity of the model structure, location will not be used as a parameter to weight the impact of model components but rather as a feature to help cluster information. A more detailed discussion is presented in section 3.5.

2.5.4 Time-series modelling approach in distribution systems

There are relevant challenges in increasing observability and controllability in the distribution system to achieve this goal. Therefore, the system identification approach has shown a relevant impact on getting the dynamic nature (including the static or zero dynamics nature) of the distribution system based on measured data. However, it is important to highlight that the distribution system with high integration of renewables is constantly changing. Even traditional systems are non-static in the sense of load consumption [160, 161]. The voltage and power in the power system are constantly changing according to the consumption patterns and generation profile, which will depend on the type of DGs introduced into the

system [14, 162, 163]. Even the change on measured data must be considered in the model construction.

From a modelling perspective, traditional voltage analysis and control in the power system have been carried out considering worst-case scenarios, i.e. the maximum load demand with the minimum generation of power injected from renewable energy sources, or minimum load demand with maximum generation when power is exported from the distribution network. For some assets, such as small transformers or loads, there are no measurement devices installed, and their operational conditions are assumed to be at rated values. Therefore, the rated capacities of transformers, feeders, and lines are achieved more quickly and give no room for integrating any other device in the system, including the renewable energy source [69, 164]. This assumption provides a security margin in the operational conditions of the distribution system, even if it does not represent the reality in most cases. For example, maximum energy consumption is not achieved during the period of maximum generation of PV units. However, the assumption of stressed conditions represents a limitation on the number of renewable units integrated into the feeder, as the voltage easily surpasses the allowed voltage range once the units start to operate [68].

One solution for modelling these variables is to consider the time-series variable as random variables and model them using traditional parametric or non-parametric modelling. Some approaches have been presented in [163, 165–169]. This provides a good approximation, but its use for voltage analysis is highly limited to obtain a good linear approximation. A more accurate approach involves running the power flow calculations with the actual power profiles for both loads and generators. Therefore, all profiles should include an associated uncertainty since profiles are changing over time. This is now possible with the measurements installed in the distribution system and the computational power of several power system simulators [47, 160].

The paradigm of using time-series measurements of system variables to construct models for control has seen widespread use and validation in industrial applications outside of the power domain, stretching back several decades [15–18, 170, 171]. While time-series analysis has found use in power distribution network modelling and analysis, particularly in support of power flow analysis considering the presence of renewables [161], topology detection [98, 172, 173], and reactive power control [174], to the author's knowledge no attempt has been made to identify the *broad* dynamics of a distribution system from limited available time-series data, considering the effects and time-evolution of uncontrollable exogenous variables. These exogenous variables could include consumer beha-

viour, weather predictions, the location of the generation unit, and other elements that are not part of the distribution system but interact with it, impacting the way the voltage evolves over time.

Therefore, if is not indicated something different, the term "dynamics" differs subtly from the traditional notion of power system dynamics since the relevant characteristics are not limited to just those of traditional assets, i.e., controllable power sources, but are expected to capture the dynamic effects of loads and renewable inputs on system variables such as voltage. The identification of these quasi-dynamics could be of valuable use in developing voltage control approaches for distribution systems with a high penetration of renewables but limited availability of measurements. This is also a motivation for this thesis.

The identification of these quasi-dynamics using time-series analysis is presented in [23, 24, 33, 175–177]. These works based their analysis on previous knowledge of the distribution system (i.e. system topology, pattern of consumption/-generation, etc.), to build the time-series results, but they were not focused on using measurable data to build a model that can or cannot have any physical representation, as is expected for this thesis.

2.5.5 Integrating uncertainty analysis in distribution systems

A crucial point to consider when dealing with time-series data is the uncertainty associated with the nature of the random variable that represents the data, as highlighted in the related literature. In planning and control contexts, worst-case scenarios are often relied upon (e.g. in "security constrained optimal power flow" formulations) to design operating points. However, an alternative approach is to integrate the random variables representing the data using uncertainty analysis. This method provides relaxed conditions based on statistics, rather than assuming the worst-case scenario, and thus enables the potential for increased integration of renewables based on statistical analysis. For example, in an Austrian case study, [178] presented a way to increase the hosting capacity of photovoltaic generation. Therefore, due to the significance of stochastic generation on improving renewable integration in the distribution system, its impact should be included in the analysis. Moving to probabilistic models makes decisions regarding risk levels explicit rather than implicit assumptions [179].

The quasi-dynamics that need to be modelled are nonlinear, stochastic, and multi-period, which will have an impact on the control design. This thesis aims to consider this aspect when modelling voltage, taking into account that inputs and outputs may have associated uncertainties, rather than assuming them as disturbances, which is typical in traditional approaches. At a basic level, the model

structure is expected to be similar to that presented in Equation (2.17). However, in this case, $x(t)$ represents the states, $u(t)$ represents the set of control inputs, and $w(t)$ represents the set of uncertain variables. The uncertain variable includes the measurement noise and the traditional system perturbations, such as consumption/generation profiles.

The uncertainty analysis provides information about $w(t)$, including state limits, probability density functions or cumulative distribution functions, and more. It is important to consider these factors when designing the control system. The controlled system's analysis and evaluation can also take these into account to produce performance statistics (e.g. cumulative distribution functions of the controlled system's voltages). Therefore, it is essential to produce more than just uncertainty bounds to avoid adopting the traditional conservative approach.

The uncertainty associated with load and renewable-based generation can be modelled using either set-theoretical or probabilistic approaches. In the set-theoretical approach, uncertain variations can be seen as forecast errors, where a nominal forecast is bounded by a confidence level [164, 180, 181]. Alternatively, in the probabilistic approach, profiles can be modelled as random variables, known as probabilistic power flow analysis [182]. The results of this approach are also represented as random variables [66].

The probabilistic approach can be categorised based on how uncertainties are integrated into the deterministic model base [183–200]. Monte Carlo approaches are typically employed [198], where random variables associated with load and generation can be easily integrated. These methods are particularly useful when mathematical and physical problems are challenging or impossible to model, especially when the probability distribution of the generation from random variables is required as inputs. Monte Carlo simulations using simple random sampling can provide the most accurate stochastic behaviour of the target random variables. One significant advantage of this approach is its flexibility [187].

Although sampling techniques can be complex, the Monte Carlo method utilises a deterministic model that establishes links between the variables being analysed and the uncertain input variables, and then calculates a set of Monte Carlo simulation samples of random inputs without requiring any reformulation. This is equivalent to performing a deterministic approach multiple times using different input combinations. Consequently, the same nonlinear form of traditional load flow calculations can be used for this probabilistic approach analysis. However, achieving convergence requires a significant computational effort.

Other methods for incorporating uncertainties into the probabilistic power flow approach are the analytical methods, also known as approximate techniques

based on Linearization [183, 185, 189–192], or by using cumulant calculations combined with the Gram-Charlier expansion theory [184, 186, 188, 193–195]. These methods are based on convolution techniques for solving calculations with probability density functions of stochastic input variables. The resulting probability density functions represent the random variables of line power flows and system states. Assumptions are made in these methods, such as the linearisation of load flow equations, the assumption of a normal distribution for the load (although this is not a mandatory condition), the assumption that the probability distribution functions of random variables are known, and the assumption of independent or linear-correlated power variables, as well as a discrete distribution for generation.

The majority of the analytical methods outlined are restricted to linearisation of the initial model, which results in fast computation but imprecise results. As the load flow equations are non-linear, and the input power variables at different nodes are typically not entirely independent or linearly correlated, this poses a challenge when solving probabilistic approaches through this method. Consequently, the primary drawback of this approach is that it linearises the power system, which overlooks certain dynamics such as OLTCs behaviour. The approximate methods take into account only a set of deterministic outputs from the original model to obtain the probabilistic results. First, the deterministic load flow problem is solved at various sample points, and then each result is assigned a weight to estimate the output moments. This mechanism is quite similar to Monte Carlo simulation, but fewer samples are needed, and extra treatment on the results is required. Methods such as the point estimation method [201–203] or the unscented transformation [204] can be used to apply this approach. In [205], a theory based on Dimension-Adaptive Sparse Grid Interpolation and its combination with Copula is used to obtain the uncertainty analysis. In [206, 207], a method called the Common Rank Approximation (CRA) method is proposed. This method relies on a simplified system coupled with a rank-comparing process and provides high accuracy, low error values, and considerable time savings.

A typical limitation of these approaches is that they do not yield the probability density functions of outputs; rather, they provide their statistical moments. The pace of the probabilistic analysis relies on the number of uncertainties. Despite this, approximate methods strike a balance between Monte Carlo simulation accuracy and analytical method speed, rendering them a viable choice for high-dimensional uncertainty quantification.

To summarise, with the introduction of stochastic wind and solar generations, deterministic models are no longer sufficient to understand the power system.

Several problems in this area are nonlinear, stochastic, and multi-period. By moving to probabilistic models, decisions regarding risk levels become an explicit decision rather than an implicit assumption [179]. For control action, it is not enough to consider uncertainties only as a bound that stresses the control requirements, but it must offer an opportunity to provide flexibility in the control targets based on statistical analysis. This section demonstrates the challenges in integrating uncertainties for power system studies, particularly for the voltage control problem. There is a significant trade-off between complexity and calculation speed. A comprehensive understanding of the time-series calculations involved in the dynamics related to the voltage in distribution systems is required to determine the appropriate approach for modelling uncertainty associated with load and generation.

2.5.6 Selection of proposed model structures

With the revision presented above, the requirements for a potential model structure can be summarised as follows:

- The model should be able to capture the relevant quasi-dynamics of the distribution system, including patterns of load consumption and generation injection.
- Measurable data should be in the form of time-series vectors and should be used as input to produce models, taking into account the uncertainties associated with the selected inputs.
- The model structure should start from a case of no previous knowledge of the distribution system and build a purely data-driven model.

State-space representations offer a promising solution to meet these requirements since they provide a better understanding of the quasi-dynamics, which can aid in the development of a tool for predicting (and potentially controlling) voltage. Spatial characteristics need to be detected (though not necessarily modelled), and therefore, any sense of location would be included as an input feature instead of being included in the model structure.

Based on this, several structures can potentially be explored to produce the model. One such structure is the linear autoregressive model, such as ARX and ARMAX, which has been used in this context and can be easily represented in state-space form. Additionally, Koopman operator-based representations show potential as a tool for directly obtaining a reduced-order representation that is easy and compact to implement, e.g., DMD. These can be contrasted with more

traditional tools such as the subspace-based methods (a combination of ERA and OKID algorithms), which has been used to identify systems based on measurable data due to their capability of decompose measurable data in equivalent "train of impulses" that can be used later to identify an equivalent system. A detailed discussion of their characteristics and application in this thesis is presented in Chapter 4. None of these models were used to provide a representation of complete distribution systems in the works presented. Normally, a Single-Input and Single-Output (SISO) representation was used to simplify the analysis, which also represents a good opportunity to explore in this thesis (with either MISO or MIMO representation).

2.6 Conclusions

This chapter has covered a significant amount of information related to the research conducted in this thesis. Firstly, a background on the subject of voltage stability in the context of distribution systems was provided. Sections 2.1 and 2.2 presented the main voltage-power equations for distribution systems with non-lossless conditions in a reduced two-nodes system to explain the relationship between both variables. After reviewing possible scenarios, it was concluded that this problem does not correspond to the classical voltage stability problem, and the matter of interest in this case was far from any critical point in the curve that represents the voltage-power relationship. Consequently, the model to be considered can be approximated in a linear representation without losing precision or the sense of the actual operational conditions.

The addition of uncertainties and exogenous variables was presented in Section 2.3, in which the possibility of using MPC to improve the operation of the distribution system and create more opportunities for integrating renewable generation units was discussed. A key takeaway from this section is that in order to achieve this voltage control objective, it is necessary to maximise the renewable injection into the distribution system while considering the operational restrictions (voltage limits, lines and transformer loadability, etc). In this regard, MPC was considered as a potential tool for achieving this goal. Therefore, a model that can provide a good representation of the current network status is required, in which the main quasi-dynamics are captured to improve the classical rule-based approach. This traditional method is more focused on voltage restrictions and does not provide much scope for optimising the integration of renewable energy sources.

The problem of observability and controllability in distribution systems was

introduced in Section 2.4. Subsequently, in Section 2.5, the problem of modelling time-series data in power system applications was introduced, along with the requirements of new metrics that incorporate time into the modelling approach. Traditional metrics primarily focus on producing indices based on prior knowledge of the steady-state condition of the system, particularly its topology configuration. This can significantly limit the evaluation of the variable nature of generation/-load units and the expected changes in system topology. One potential metric explored as an indicator relates to the power injected into each node and the size of its impact across the distribution system, without any prior knowledge of the system topology, which could be based on Fisher-Z transform. Another potential metric that could be included in the modelling approach is one that provides a sense of distance between measurements, as a spatial-temporal characteristic. These metrics can explain aspects of the distribution system and its assets, are easily measurable, and have the potential to become an input for the proposed modelling approach.

After this discussion, various methods for obtaining the static and dynamic characteristics of the system were explored. Additionally, it was introduced system identification as a tool for producing models that relate measurable variables to obtain a representation of the distribution system. Since it is expected to produce linear representations that can be used with time-series data and capture the quasi-dynamics associated with the random variables that represent generation and consumption, state-space representation is a potential solution to meet these requirements. While spatio-temporal system identification approaches could be a possible solution for this, for this thesis, it is expected to measure the spatial characteristics instead of modelling it. Therefore, linear autoregressive models such as ARX and ARMAX can be explored as possible structures to be implemented in their state-space form. Koopman operator-based representations such as DMD and subspace-based methods such as a combination between ERA and OKID have shown good results in different applications for obtaining reduced-order representations of various kinds of systems and can also be considered.

From this review, several observations emerge. The nature of the distribution system, which is partially observable and limited in terms of controllability, poses a challenge for producing models for control applications. These models must not only incorporate classical electric parameters but also account for uncertainties associated with exogenous variables, such as weather and spatial location, which become increasingly important with the integration of renewable generation units. Moreover, stakeholder actions are dependent on various factors, such as time of day and season of the year. New metrics are required to describe distribution

systems without relying on any prior knowledge of classical electric parameters and that reflect the actual status of the network based solely on measurable data. Such metrics will facilitate the production of models that can be easily integrated with MPC approaches, and the control objective can be adjusted according to real conditions and operations. These issues represent interesting research areas that will be addressed in the remainder of this thesis, within the context of voltage control in future distribution power networks.

Chapter 3

Data-Driven Characterisation of Distribution Systems

3.1 Introduction

Data-driven approaches are becoming a new trend for system modelling in voltage control applications at both transmission and distribution systems level. The high penetration of renewables and a better understanding of customers require flexible schemes that adapt according to the system's reality. This task is more challenging in distribution systems because of the limited observability, and most methods rely on classical non-scalable physical-based models [19, 93, 96].

Many data-driven approaches to power system modelling and estimation assume that the system is fully observable, implying that the measured data reflects the time evolution of the modes to be analysed [88, 102]. Correspondingly, several power system control approaches assume full controllability, meaning the manipulable inputs in the system are able to excite all modes of interest [89, 90, 103, 104]. In this thesis, a more realistic perspective is adopted: considering limited knowledge of the system topology and parameters, and limited availability of measurements, what information can be extracted regarding the controllability and observability of the distribution network? Which nodes are most useful for observation and control, in the context of managing the system voltage?

In this thesis, the foundations for a new approach are proposed through an investigation and analysis of the relationship between certain measured data in the network and what can be inferred about the spatial-temporal profile of the system voltage. The particular aim is to develop and validate metrics that characterise and quantify this dependency, allowing the wider impact of nodal power fluctuations (owing to changes in load or power injections) to be predicted. An

unbalanced network with an arbitrary level of penetration from renewable power sources is considered, and nothing in particular is assumed about its topology and parameters.

For the network used as a reference (IEEE 123-nodes), OLTCs are assumed to be in operation to control voltage locally. This is also compared with the addition of capacitor banks into the operation of this network. Additionally, some scenarios with a meshed system are considered to contrast the response of this approach in both radial and meshed configurations. This will illustrate the conditions in which the metrics are calculated and provide relevant information about the system, even with only measured data available. This approach can be extended to networks with many other kinds of devices/control strategies, or even different times of the day and seasons of the year, as the approach relies solely on measurements of voltages and powers, which can capture the nature of the random variables that describe these operational conditions. This chapter discusses how the metrics are able to detect these conditions and characteristics of the system without the full knowledge of the topology.

The examination for this approach focuses on how measurements of power flows through lines provide information on the equivalent impedance of the system. The thesis goes on to propose matrices that use the Fisher z-transformation of the Pearson correlation values between measured nodal voltages to indicate how a voltage variation caused by a nodal perturbation propagates through the network, in terms of which other nodes will also see perturbations. Building on this, the thesis proposes to *quantify* the voltage variation and propagation in response to a perturbation event, in the absence of impedance measurements or knowledge, using the covariances of voltage measurements at selected nodes. These developments support the wider aim of developing a framework for identifying distribution system dynamics under limited knowledge and measurements by enabling the identification of the key nodal voltages in the system and offering a non-parametric characterisation of how they respond to inputs.

Once the proposed metrics are constructed from time-series nodal voltage and power injection data, it is indicated how they can provide information on two main aspects: (i) how perturbations or control actions propagate through the distribution network and (ii) how electrically close nodes are, and how similar and correlated their voltages will be. The potential use of these new metrics is discussed for obtaining spatial-temporal characteristics of distribution systems, and for identifying suitable variables of interest and candidates for inclusion in a reduced-order model of the system for voltage control purposes.

The metrics are computed from time-series measurements of system data, in-

cluding power injections into lines and nodal voltage in nodes, obviating the need for a physics-based model and parameter estimation. The effectiveness and validity of the proposed metrics are evaluated through case study simulations on a 123-node test network subject to diverse types of perturbations. Furthermore, the model-free, data-driven approach paves the way for capturing the effects of difficult-to-model, exogenous variables, such as renewable power injections and load profiles. Additionally, an algorithm for the modelling based on the available data is proposed. The major contributions of this chapter are:

1. It is proposed a *data-driven* approach to characterise distribution systems based only on time-series measurements that provides relevant spatial-temporal information of the distribution system and the perturbations that occurs during operation;

2. It is proposed a methodology to analyse and reduce model complexity of distribution system based on the electrical distance by using Fisher z-transform and voltage covariance from measured data;

3. It is demonstrated the use and investigate the efficacy of the metrics proposed via simulations on an unbalanced distribution system (IEEE 123-node test network) under various scenarios; and

4. It is discussed the interpretation of the new metrics for the spatial-temporal description of the distribution systems and potential use in developing models in control applications.

3.2 General problem statement

In Section 2.1, the stability problem was introduced to evaluate the type of dynamic system that can potentially be obtained in the system representation. After further discussing the integration of exogenous variables in this system explained in time-series data, the quasi-dynamic system model should consider more components than just the electrical parameters. Therefore, this research considers a general distribution network composed of a set of nodes $\mathcal{N} := 1, \ldots, N$, where the voltage–power quasi-dynamics are assumed to be governed by Differential–Algebraic Equations (DAEs):

$$\dot{x} = f(x, u, w, t), \tag{3.1a}$$

$$y = g(x, u, r, t). \tag{3.1b}$$

As stated, these DAEs are sufficiently general to capture all quasi-dynamics in the system—including electromagnetic phenomena—but for the purpose of this thesis, they are assumed to reflect the timescale of interest for voltage control (i.e. in the range of seconds to minutes). Therefore, these DAEs capture the voltage–power physics of the electrical network and the non-deterministic, time-varying and quasi-dynamic actions of consumers and producers. In equation (3.1), t denotes time, x is a vector of (internal) states, whose time evolution is described by equation (3.1a). u is a vector of control input variables, including nodal active and reactive power injections, y is a vector of measured outputs, collecting the voltages observed at certain nodes, and w and r are vectors of uncertain and/or exogenous variables affecting, respectively, the state evolution and system output.

The motivation for this thesis is to consider the desire to perform voltage control in the distribution network, where the quasi-dynamics (3.1) are unknown. Moreover, the system comprises a large number of unmeasured states, and only a subset of nodal voltages may be measured, along with a subset of nodal power injections available for manipulation for control purposes. Therefore, a *system identification* process is required to construct a simple yet sufficiently accurate model of the power–voltage quasi-dynamics at the timescale of interest for voltage regulation. Prior to identification, it is necessary to determine *which* nodal voltages should be measured and which nodal power injections should be made available for control so that (i) the measured voltages provide an adequate picture of the voltage profile across the network, and (ii) the nodal power injections made are capable of adequately influencing the voltage profile across the network. In other words, which nodes in the system are *critical* to the observability and controllability of the voltage? The aim of this chapter is to answer this question by proposing a set of metrics that utilise real system measurements to assist in determining the critical nodes for observability and controllability.

It is considered an unbalanced radial network composed of N nodes. The topology and parameters of the system are not completely known, and neither is the composition and nature of the loads. Renewable generation is present in the network in the form of uncontrollable power injections at certain nodes. It is assumed that measurements of voltages are available at the feeder and some nodes in the network, but it may not be known exactly where these nodes are relative to other nodes. Figure 3.1 illustrates a typical scenario under such assumptions.

A trio of nodes, n_k, n_j and n_e, are shown in relation to the feeder node n_i; it is known that n_k and n_j are connected by a line, but their specific connections with upstream and downstream nodes are not known beyond that there exists a path to the feeder and a path to downstream nodes. The parameters of the line

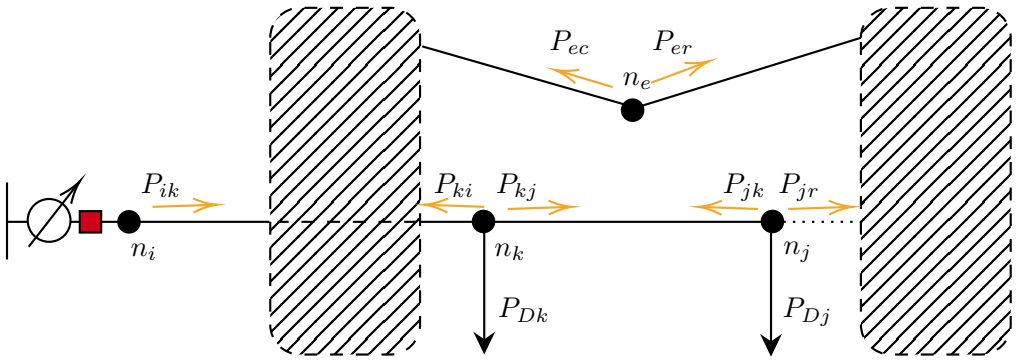


Figure 3.1: Illustration of a distribution system with partially known topology and connectivity

between n_k and n_j are not necessarily known. It is known that node n_e is within the geographical vicinity of n_k and n_j but its precise connectivity is unknown. The OLTC that controls the voltage at the beginning of the feeder in this illustration will respond to regulate the voltage and maintain it within the operational limits (if activated, which is commonly the case). This can respond locally by sensing the voltage at the beginning of the feeder, even if it is not able to measure the voltage in all nodes downstream. There could be more devices, such as other small OLTCs or capacitor banks in the area, that are not available for measurement and cannot be controlled. Nevertheless, the total power flowing through the lines will be altered up to the beginning of the feeder, and the response of this voltage control will correspond to this total power balanced across the system. Therefore, if nodal voltages are measurable and power injections and flows are known at such nodes and other such nodes in the network, what picture can be gleaned of the overall network voltage profile in response to power injections—either controllable or uncontrollable—and demands?

3.3 Analysis and synthetic production of time-series data

The first task this chapter corresponds to analyse features that have a high impact in the voltage of the distribution system. Therefore, it is required to start exploring from a high-fidelity model which technical aspects, nodes, variables or scenarios affect the operational voltage. As part of this thesis, a reference data set is required to carry all the corresponding project. Synthetic data must be constructed from a software that allows modelling the distribution system with its components and their corresponding control functions (for instance, inverter control functions, OLTC and regulator controllers, etc.). For this thesis problem, only PV units will be considered as source of renewable energy in the distribution system. Therefore,

uncertainty associated with this renewable energy source will be studied.

According to [68], the reference data must consider the following characteristics:

- Time-series simulations must be conducted. Therefore, several load flow calculations are required to analyse the profile instead of the traditional steady-state analysis with fixed values. This involves synthesising time-series data via a sequence of load flows with different parameter snapshots (e.g., load). It should not be confused with a full dynamic simulation, which models the time-varying behaviour of a system described by ordinary differential equations or partial differential equations.

- The distribution system could be balanced or unbalanced. A "balanced" system is one where all line voltages are equal on each phase, and therefore all line currents are also equal. An "unbalanced" distribution system is usually composed of unsymmetrical loads, which implies that voltages and currents are not the same on each phase. Simulation software must be able to work for both scenarios if required.

- Variability associated with load and generation must also be integrated. Therefore, the deterministic approach must be changed into a probabilistic approach.

A combination of OpenDSS and MATLAB was implemented to deal with all of these requirements. OpenDSS is a free licensed software developed by Electric Power Research Institute (EPRI) [208], which focuses on modelling distribution systems and all of the devices and components that can be connected at this part of the power system (including voltage regulators, capacitor banks, inverters, among others). This software also allows steady-state simulations, dynamic domain simulations and fault-events simulations. Figure 3.2 illustrates the integration of both through a COM server developed for OpenDSS to communicate with MATLAB. The latter is relevant for developing scripts that automate some calculations, including the selection of random inputs (applying Monte-Carlo approaches) and modelling some control actions and functions if required.

3.3.1 Modelling of the distribution system

The simulations depicted in this thesis consider the IEEE 123-node test network, an unbalanced distribution system with a total of 269 nodes. This is shown in Figure 3.3, including the positioning of voltage regulators and the states of switches

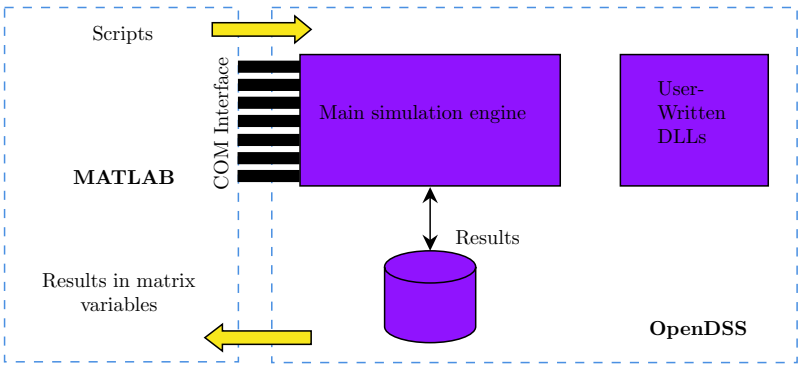


Figure 3.2: COM interface of OpenDSS with MATLAB

in the network (where green denotes an open switch and red a closed switch). It will provide a baseline dataset for the system in the absence of control and exposing the cause-effect relation between renewables penetration and system voltage without any obfuscation by the control.

3.3.2 Modelling of load profiles and daily solar radiation

For implementing time-series simulations, the CREST Demand Model from McKenna and Thomson [209] was used for both solar radiation and residential load profiles. All data obtained from this model have a resolution of one minute, resulting in 1440 load flows for simulating the profile of one day. The model can consider all calendar days of the year.

3.3.2.1 Generation of load consumption profiles

The consumption profiles were created first, and the model used can consider between 1 to 5 people in a house, distinguishing between weekdays and weekends. First, the algorithm scanned all the loads connected in the distribution system (using a constant power model). For the distribution system used as a reference, there are some nodes that can have single-phase loads and others with three-phase loads (more details in Appendix E). For each node, a consumption profile is assigned by selecting a random number of occupants in the house (between 1 and 5, and only residential consumers are considered), obtained from a MATLAB random generator function. This number is then input into the CREST Demand Model, which randomly assigns the type of devices and consumption profiles for each case, taking into account the season and time of day (previously defined). This produces a daily consumption profile ranging from 0 to 1. Therefore, the profile is scaled up by the individual household demands of the rated power of the load over each

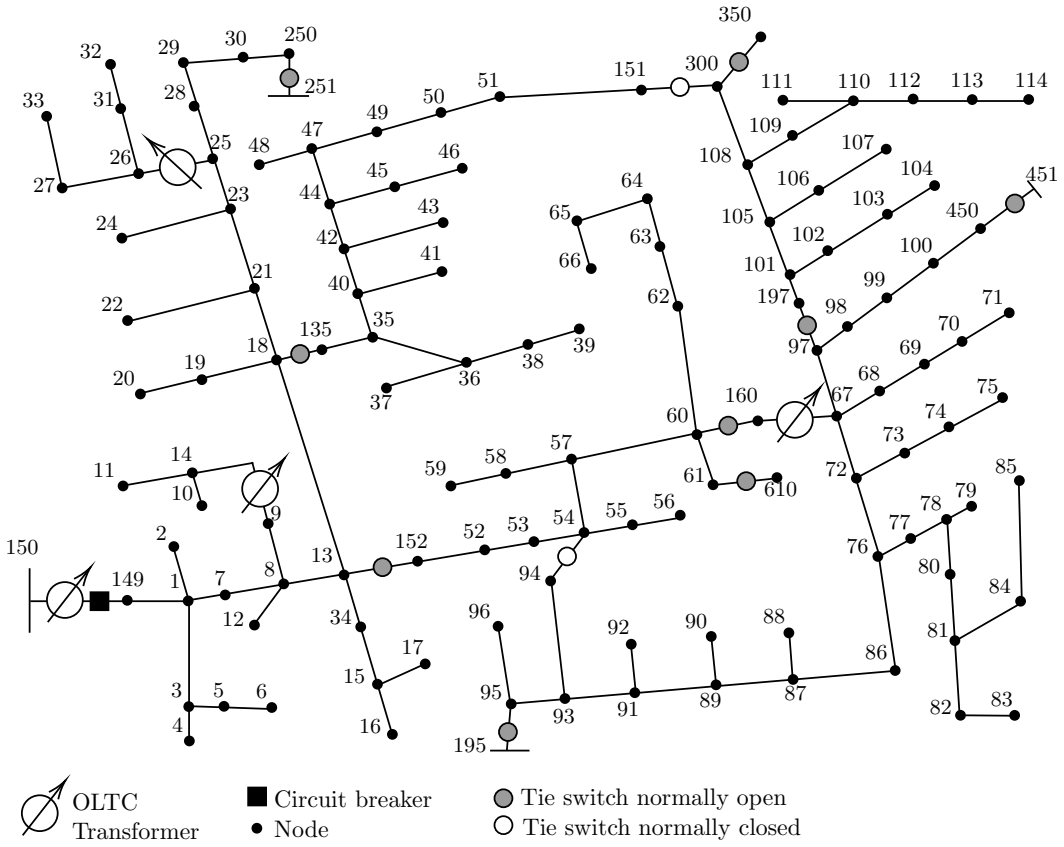


Figure 3.3: IEEE 123-node unbalanced distribution system

node. This represents the assumption in which each transformer over each node groups houses with the same number of people and the same consumption behaviour. Therefore, profiles were generated for each possible number of people in the house and for each type of day, resulting in 5x2=10 possible profiles. The unbalanced condition of the system will be reflected in the nature of having different profiles in different nodes/phases. Figure 3.4 shows some examples of individual load profiles and Figure 3.5 shows one day example of the total power profile at the main feeder when there is a 30% of renewable generation integrated into the system, to illustrate the unbalance condition of the system.

3.3.2.2 Generation of solar radiation profiles

Once the load consumption profiles had been selected and integrated, the solar radiation profiles for each day of the year were created. These profiles were defined based on the geographical location of the distribution system and the cloudiness index denoted as B_{index}. The solar profiles for this model were defined according to Equation 3.2:

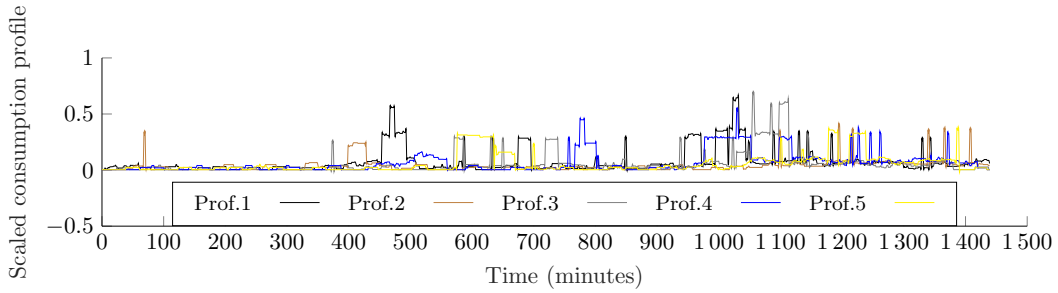
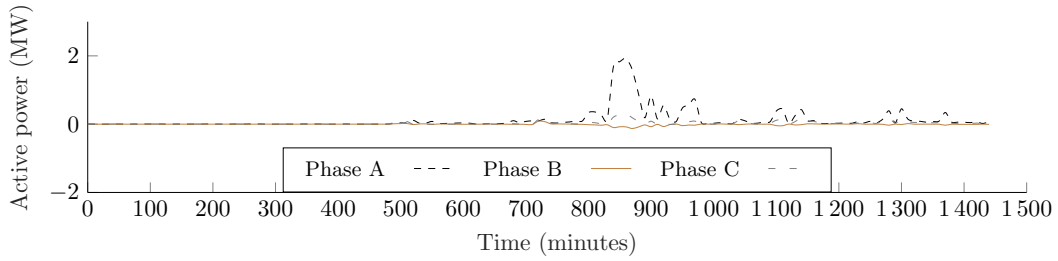
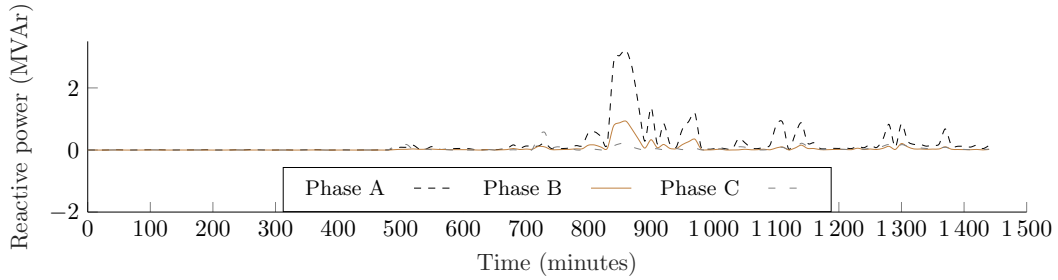


Figure 3.4: Examples of profiles obtained after using CREST Demand Model and the total power flowing through a representative three-phase node



(a) Active power profile



(b) Reactive power profile

Figure 3.5: An example of a day total power profile at node S149 (main feeder)

$$PV_{prof} = \eta_{panel} * \eta_{inv} * A_{panel} * H_{index} * B_{index} \tag{3.2}$$

where η_{panel} and η_{inv} are the panel and inverter efficiencies, respectively. The A_{panel} area of the panel and the H_{index} irradiance profile of the corresponding month were used to create the solar profiles for each day of the year. Other relevant assumptions are shown in Table 3.1.

The irradiance profiles were generated using the fixed parameters mentioned earlier (assuming the location of the distribution system is Sheffield, with the same dimensions of solar panels and efficiency coefficients for each generation unit), and the cloudiness index was selected randomly from a matrix of 100 possible

Variable	Value
Latitude	53.38°
Longitude	-1.47°
Day of the year that summer time starts	87
Day of the year that summer time ends	304
Slope of panel	35°
Azimuth of panel	0
Panel area A_{panel}	10 m^2
System efficiency ($eta_{inv} * eta_{panel}$)	0.1

Table 3.1: Summary of panels and inverters characteristics.

daily sky brightness/cloudiness index values stored in the CREST Demand Model. This value was obtained using the MATLAB random generator. For each type of day (weekday/weekend, summer/winter) and each cloudiness index, the model provided a solar radiation profile in a similar way to the load profiles described in the previous section, with values ranging from 0 to 1 for a typical day. These values were obtained for each day and scaled to the rated power of the generation unit.

The way in which PV units are integrated was determined based on the amount of power that was integrated into the system. This is expressed as a percentage of the total rated load power installed in all nodes. For example, a penetration level of 50% means that a total amount of 50% of the equivalent rated power of all loads installed into the distribution system was integrated. The size of each unit was randomly selected using the MATLAB generator function, and commercial values for generation units were used, ensuring that all units could achieve the previously defined power level. The power injection was either single-phase or three-phase, depending on the load installed at that node. Then, all units share the same generation profile, assuming that all of them are close enough to receive the same amount of solar irradiance.

The location of each PV unit is a random position within all nodes that have loads on the distribution system. To achieve that, all nodes from the distribution system model were scanned, reduced only to those with load connected, and assumed that the possibility of installing a PV unit was the same on each node. A more specific discussion of this is presented in next section.

3.3.3 Modelling of uncertainties

The next step was to identify uncertainties associated with each component. Typically, the distribution system topology and consumer-rated power are assumed to be fixed since the system already exists. It is assumed that the IEEE 123-nodes system will maintain the same configuration topology (no reconfiguration) and the rated power of each load, as load movement is not expected. The change will only be focused on the load profile each day. Therefore, the location and rated power of the loads were known beforehand.

The locations, daily profiles, and rated power of the solar panels were randomly selected. The location of each PV unit was determined using a uniform random variable, meaning that all nodes connected to loads were equally likely to receive a PV unit. The rated power of the inverters was defined using the MATLAB function 'random' within a range between the minimum and maximum commercially available values, but not exceeding the maximum load installed in the distribution system. The random variables associated with the variability of solar profiles were selected from the previously mentioned pool of profiles, and the chosen curves were scaled to the rated power of each PV unit. Similarly, the load profiles were defined based on the variability integrated in the model of McKenna and Thomson [209], and the profiles were scaled according to the rated power of each load.

3.3.4 Flow chart for calculation algorithm

The planned flowchart for running all simulations using OpenDSS and MATLAB is shown in Figure 3.6. It shows how random variables for both load and generation are assigned to each element, following the explanation given in Section 3.3.3. The amount of PV rated power integrated into the system is determined based on the renewable penetration level, which represents the amount of renewable rated power installed in the distribution system as a percentage of the rated power of all loads combined. After the locations, rated powers, and profiles for each component were defined, load flow was run in OpenDSS for a time-series data set of one day, consisting of 1440 load flows. The results were stored in MATLAB variables and saved for further analysis. In this case, voltage profiles of 1440 points for each node were obtained, representing the voltage value every minute. The entire process was repeated for each new scenario simulated, with only the topology of the distribution system and the location and rated power of the load being fixed, while the rest of the elements and their uncertainties were redefined for each simulation.

Table 3.2: Summary of simulation cases

	Summer profile: 217 days for possible combinations Each day: 1 minute resolution profile 1440 load flows	Winter profile: 148 days for possible combinations Each day: 1 minute resolution profile 1440 load flows
1000 possible scenarios (weekdays) Changing rated power, position and profile of PV unit.	1000 possible scenarios 1440000 load flows (Summer - weekdays)	1000 possible scenarios 1440000 load flows (Winter - weekdays)
1000 possible scenarios (weekends) Changing rated power, position and profile of PV unit.	1000 possible scenarios 1440000 load flows (Summer - weekends)	1000 possible scenarios 1440000 load flows (Winter - weekends)

Table 3.2 shows a summary of the number of simulations that were done for this first stage. Relevant data was classified in type of day (weekdays and weekend) and season (summer or winter). For any of those, a combination of 1000 simulation was done (e.g., one thousand simulations for the distribution system during summer and considering weekdays). The complete process was repeated for different penetration levels, varying from 10% to 100%.

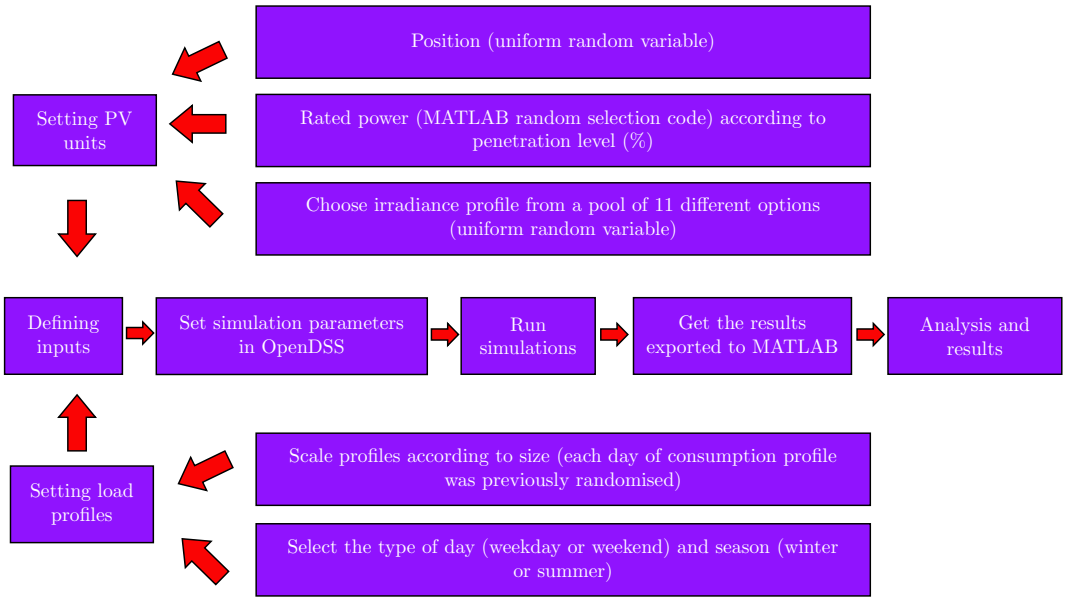


Figure 3.6: Flow chart for simulations

The results expected after running all these simulations are the Cumulative distribution function (CDF) of the voltage for each node. This procedure illustrates the probabilistic load flow using Monte-Carlo approach, which gives all details and includes non-linear dynamics associated with the distribution system.

3.3.5 Simulation results and discussion of produced data

A first observation prior to the final analysis was done for one case to illustrate the voltage variation associated with the penetration of renewables in the distribution system. Figures 3.7 and 3.10 show two nodes at two penetration levels; node 7 is chosen as one that shows small voltage variations, while node 114 is one with a large voltage variation over different scenarios. The reference IEEE 123-nodes used included some capacitor banks that are normally disconnected. To compare the response including reactive compensation, Figures 3.8 and 3.11 show the system's response with these capacitors connected. It is observed that the voltage across the system changes according to the voltage sensed by the OLTCs. However, the voltage variability for both nodes exhibits the same pattern as the penetration level increases.

As was stated in Chapter 2, it is expected as a final goal avoiding voltage issues reducing the curtailment procedure as much as possible. All countries have national standards that define allowed variations in voltages (including unbalance) in high voltage, medium voltage and and low voltage networks. For example, the standard EN-50160 for low voltage and medium voltage from the European Committee for Electrotechnical Standardization (CENELEC) [210, 211] suggests that 95% of the 10-minutes mean rms values of the supply voltage shall be within the range of +/- 10% of nominal voltage, during each period of one week. Also, all 10-minute mean rms values of the supply voltage shall be within the range of +10% and -15% of nominal voltage. Other standard is the VDE-AR-N-4105 from the Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE) [212], which suggests that the voltage range from all photovoltaic systems in the distribution grid may not exceed +/-3% in a load-free scenario. This standard is applicable only to PV generation systems, and customers in Germany are typically three-phase connected. For this thesis, operational voltage variation was not analysed according to the standard VDE-AR-N-4105. Nevertheless, the use of the voltage variation of +/-3% was used because is more restrictive for voltage variations than the suggested for EN-50160 and this generally prevents issues with excessive losses and overloading in simulations. Additionally, a potential linearisation process is easier for a narrow range of voltage variations. The ECDF for each node was constructed according to the function defined for MATLAB.

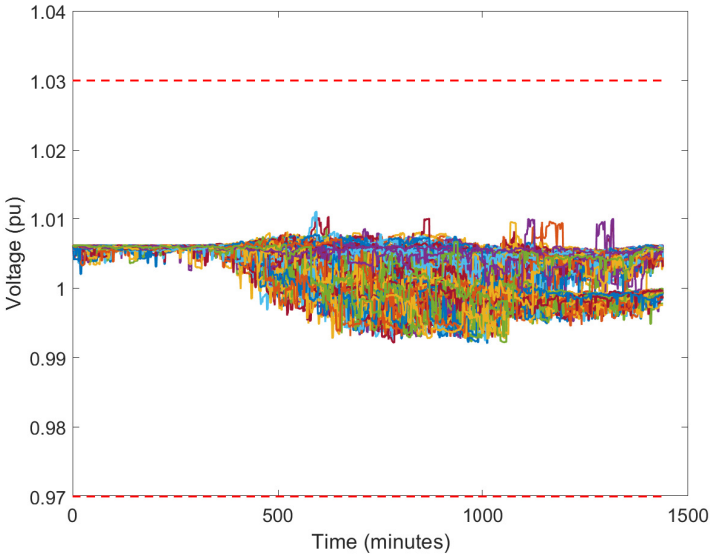
The first node corresponds to node 7, phase A (7.A). Figures 3.7a and 3.7b show the 1000 profiles obtained for this scenario with 10% and 70% respectively considering the capaitor banks disconnected. The limits were also drawn with red dashed line, in order to have a visual picture when voltage limits are violated. In a similar way, Figures 3.8a and 3.8b show the same profiles when the capacitor

banks are connected. The same information can be represented in Figure 3.9, which show the ECDF of each case, considering every single load flow for its construction. Even considering the increase of voltage fluctuation in the 70% due to the renewable energy sources, it shows a strong pattern of avoiding voltage changing above the allowed limits. In this case, the node is close to the feeder, which means that voltage regulation is done by the rest of the power system.

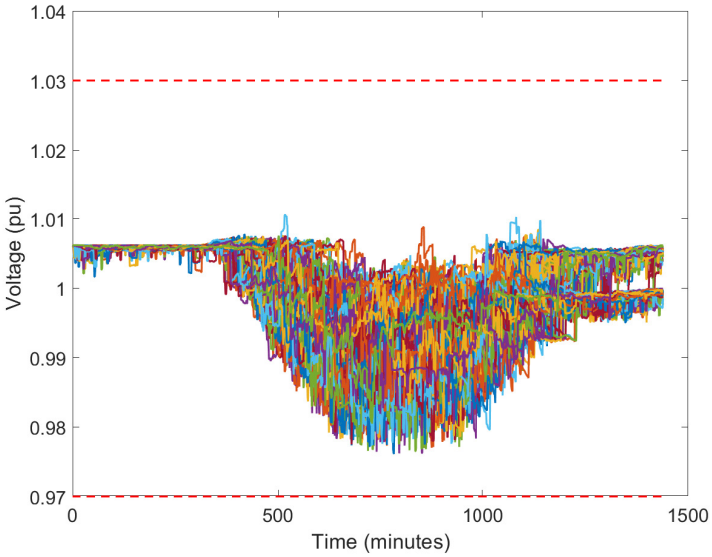
The other node corresponds to node 114, phase A (114.A). In the same way as previous, Figures 3.10a and 3.10b show the 1000 profiles obtained for this scenario with 10% and 70% respectively, considering the capacitor banks disconnected. In a similar way, Figures 3.11a and 3.11b show the 1000 profiles obtained for this scenario with 10% and 70% respectively, considering the capacitor banks connected. It is shown that the 10% of renewable penetration has some small periods of times in which voltage achieve values outside the allowed limits. Nevertheless, the corresponding ECDF curve in Figure 3.12 shows that voltage remains inside the limits in almost 99% of cases. However, it is shown for the case of 70% of renewable penetration level that there are several cases in which voltage surpasses the voltage limit. The corresponding ECDF shows that achieving the voltage values within the allowed limits corresponds to the 98% of the cases. Even if the graph shows several cases with voltage variation, the probability of having the values below the allowed limits is higher than the 95% of cases. Part of this thesis correspond to understand these variations and the impact for the control criteria. The outcome of this would result in a different criterion in which the voltage targets are achieved with a margin of probability, instead of thinking only in fixed values that reduce flexibility and does not respond to the reality in most of the cases.

Moreover, the amount of renewable energy also impacts the voltage fluctuations across the entire system. Figure 3.13 presents the average voltage profile for these 1000 simulations. It is evident that, for both nodes, the voltage variation is generally higher as the number of photovoltaic units increases throughout the system, regardless of whether the capacitor banks from the reference case are connected or disconnected.

The results of all simulations can be summarised in Figures 3.14, 3.15, 3.16, and 3.17, which provide a comprehensive overview of the results across all penetration levels and nodes. These graphs consider the number of customers that experienced voltage issues at least once, assuming one customer per node and phase. The figures utilise box plots to represent the distribution of results. Each box plot displays the interquartile range, with the bar representing the 25^{th} and 75^{th} percentiles, and a middle value indicating the median. The 'Whiskers' lines extend to the minimum and maximum values obtained from the distribution, with



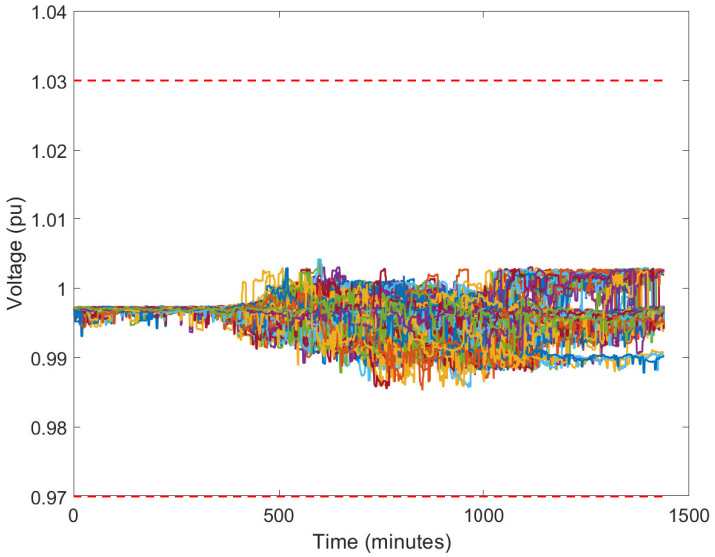
(a) Voltage profiles from 1000 simulations considering 10% of renewable penetration



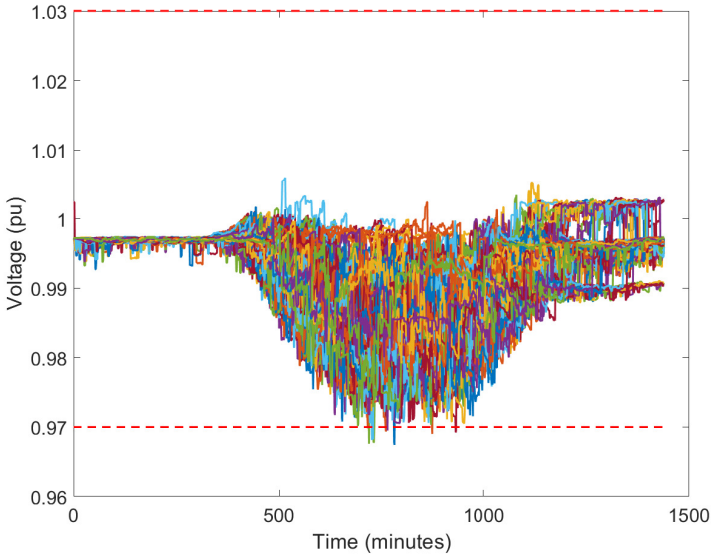
(b) Voltage profiles from 1000 simulations considering 70% of renewable penetration

Figure 3.7: Results of voltage fluctuations obtained for two different renewable penetration levels at node 7 phase A (7.A), assuming the scenarios correspond to summer season weekdays and considering the capacitor banks disconnected

any data beyond these limits considered as outliers. Additionally, for each penetration level, the figures indicate the number of customers experiencing voltages below the lower limit (or "ll", indicated by the red bars) and above the upper limit (or "ul", indicated by the blue bars). These scenarios demonstrate that there is only



(a) Voltage profiles from 1000 simulations considering 10% of renewable penetration



(b) Voltage profiles from 1000 simulations considering 70% of renewable penetration

Figure 3.8: Results of voltage fluctuations obtained for two different renewable penetration levels at node 7 phase A (7.A), assuming the scenarios correspond to summer season weekdays and considering the capacitor banks connected

a limited penetration margin without voltage issues. The figures show the median number of customers that experienced voltage outside the allowed range for the one thousand simulations. It is evident that achieving a penetration level of only 10% without significant voltage issues is possible in all cases analysed. There is

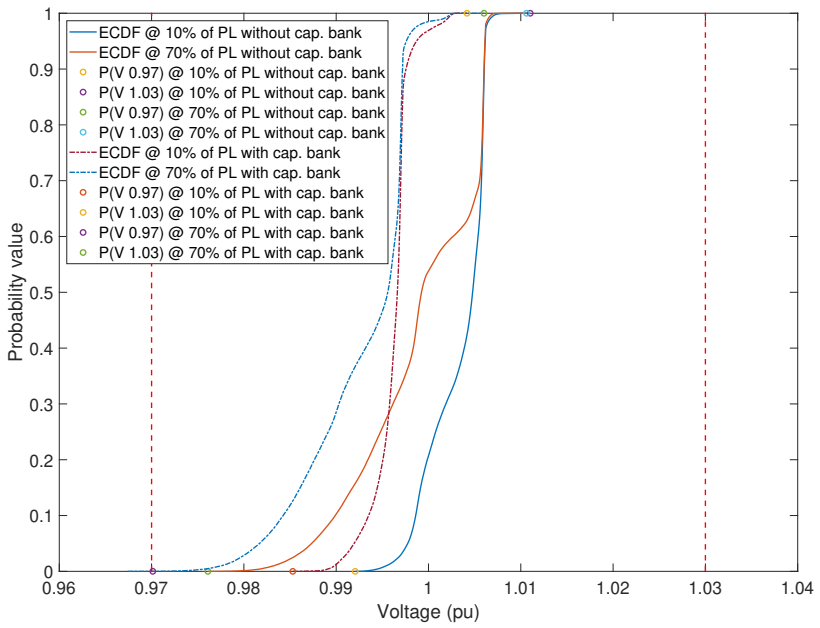
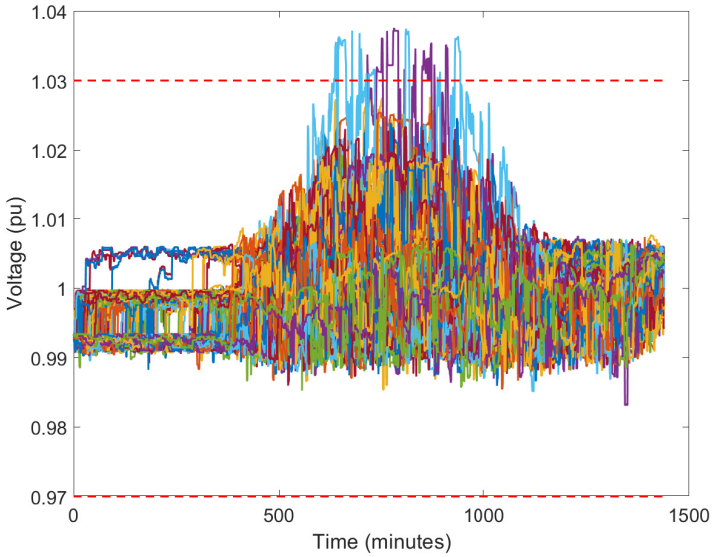


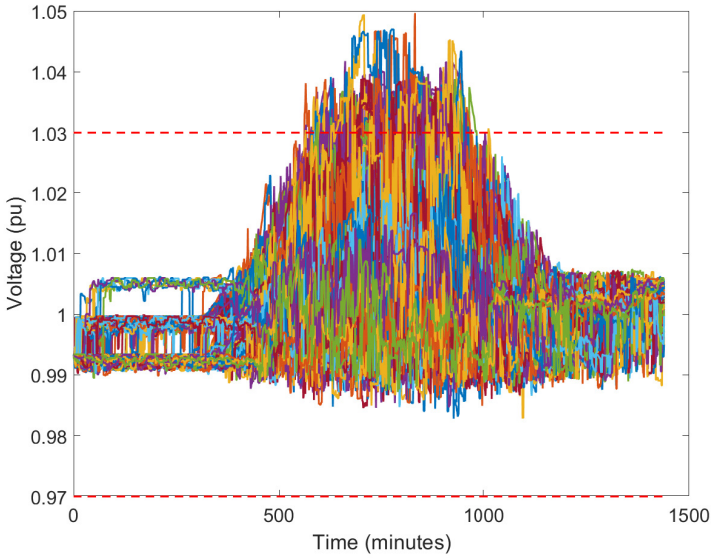
Figure 3.9: Results of ECDF voltage curves obtained for two different renewable penetration levels at node 7 phase A (7.A), assuming the scenarios correspond to summer season and weekdays

a proportional relationship between the penetration level and the number of customers with voltage issues. However, the range of variability changes depending on the type of day, season, and penetration level. It is crucial to study the properties of the distribution system carefully to identify patterns and characteristics that help reduce the number of customers experiencing voltage problems under varying conditions.

Additional circuit parameters must be analysed in order to find the parameters that helps on identifying which nodes are sensitive or robust to big voltage oscillations. These parameters should be measured on nodes in which measurement is available and should be checked as an identification parameter (for example current and voltage values, R/X ratio, among others). Moreover, to the voltage restriction, it is recommended to check current and transformer capacity during control actions. Therefore, it is required to analyse the dynamics behind the voltage operation considering fixed components and exogenous variables, to identify the relevant characteristic of the system that helps to construct the model using the system identification technique considering partial measurement of the distribution system.



(a) Voltage profiles from 1000 simulations considering 10% of renewable penetration

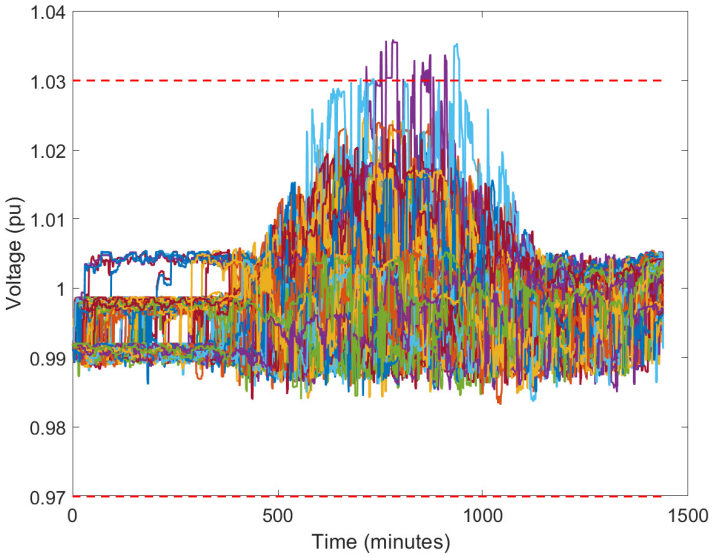


(b) Voltage profiles from 1000 simulations considering 70% of renewable penetration

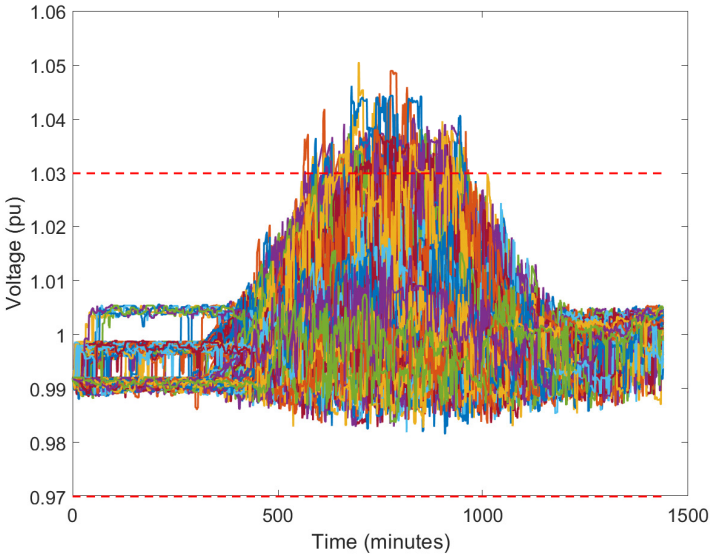
Figure 3.10: Results of voltage fluctuations obtained for two different renewable penetration levels at node 114 phase A (114.A), assuming the scenarios correspond to summer season weekdays and considering the capacitor banks disconnected

3.4 Input analysis: determining the impact of power injections

After simulations and different scenarios are run, the next step is to identify which of the network nodes are the *critical* ones with respect to providing control actions



(a) Voltage profiles from 1000 simulations considering 10% of renewable penetration



(b) Voltage profiles from 1000 simulations considering 70% of renewable penetration

Figure 3.11: Results of voltage fluctuations obtained for two different renewable penetration levels at node 114 phase A (114.A), assuming the scenarios correspond to summer season weekdays and considering the capacitor banks connected

(power injections). In this section, it is presented an analysis of how real data measurements of line power flows and nodal voltages provide, in the absence of knowledge on system topology and parameters, information on the impact of power injections on system voltages. For simplicity, it is assumed that measure-

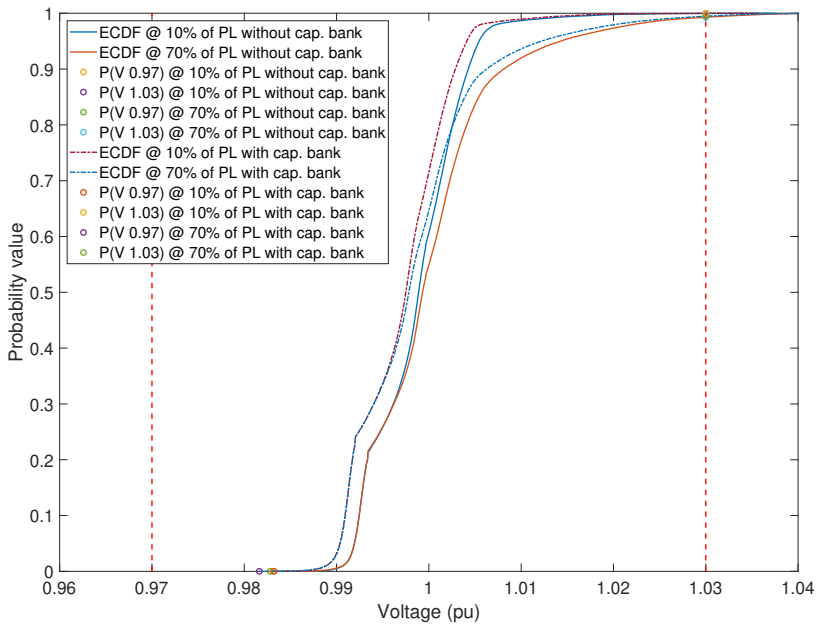


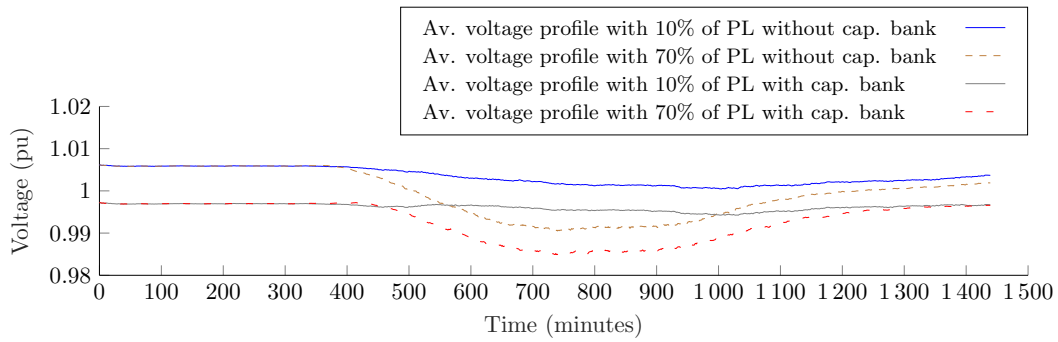
Figure 3.12: Results of ECDF voltage curves obtained for two different renewable penetration levels at node 114 phase A (114.A), assuming the scenarios correspond to summer season and weekdays

ments are noise-free. Additionally, a scenario of critical perturbation is considered since the desired condition in which data want to be explored is when voltage achieve values beyond limits. Unfortunately, as showed in Figures 3.9 and 3.12, the probability of having these operational issues is low for this system in operational conditions. Therefore, it was considered for the rest of this chapter a fixed power profile with a high perturbation size instead of a specific penetration level.

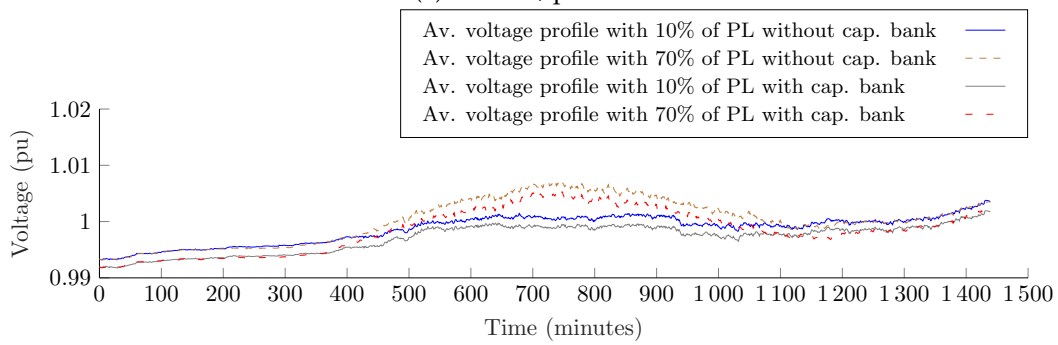
3.4.1 Definition and description of test network and scenarios

As stated previously, for this particular analysis only critical perturbations are observed using the same methodology presented in Section 3.3.3. These scenarios are critical in the sense that the load/supply perturbations are they contain lead to significant overvoltage events at some nodes at certain times during the day. Scenarios involving other nodes and smaller perturbations are not presented since they were found to cause less critical voltage responses. The following scenarios over two days were simulated in the system:

S1 Single-phase perturbation at node 85, phase C. Load consumption with a rated power of 400 kW and 200 kVAr.



(a) Node 7, phase A



(b) Node 114, phase A

Figure 3.13: Average of voltage profile for a typical summer weekday, considering two different penetration levels

S2 Single-phase perturbation at node 85, phase C. Photovoltaic generation with a rated power of 400 kVA at unity power factor.

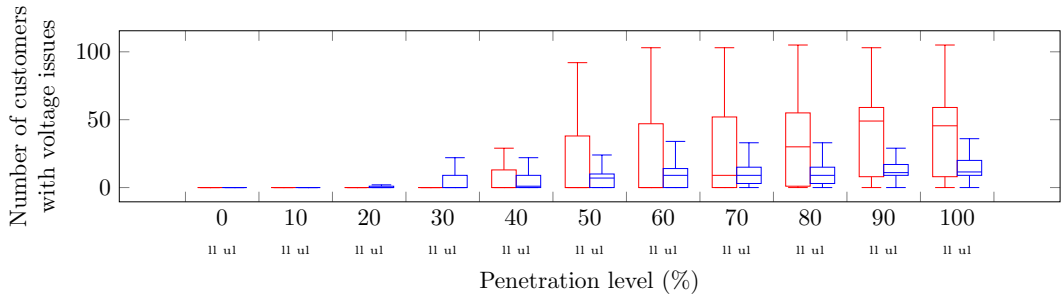
S3 Single-phase perturbation at node 66, phase C. Load consumption with a rated power of 400 kW and 200 kVAr.

S4 Two-phase perturbation at nodes 82, phase A and 85, phase C. Load consumption with a rated power of 400 kW and 200 kVAr.

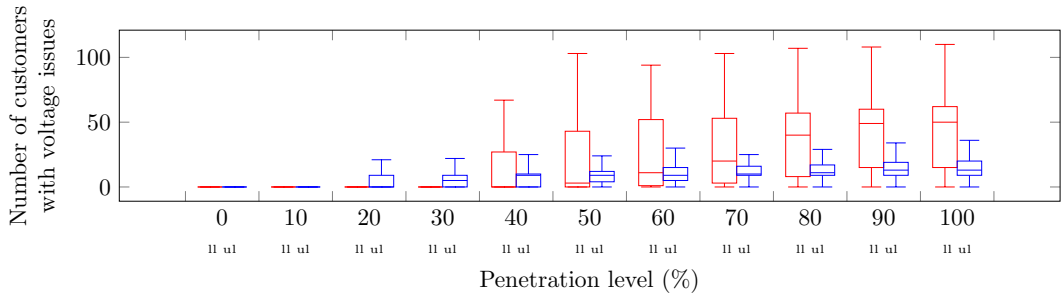
S5 Three-phase perturbation at node 48. Load consumption with a rated power of 400 kW and 200 kVAr.

S6 Single-phase perturbations at nodes 66, phase C and 85, phase C (synchronized). Load consumption with a rated power of 400 kW and 200 kVAr.

For the scenarios with high consumption, the perturbation reflected in these two days of simulation was a result of scaling the consumption profile on a critical day at the node where consumption was considerably high and commonly reflected voltage issues, using the simulations developed in the previous section. For



(a) OLTCs connected and capacitor banks disconnected

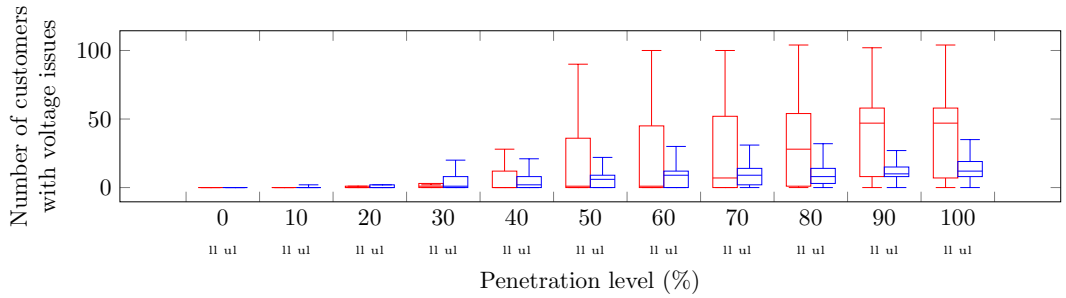


(b) OLTCs connected and capacitor banks connected

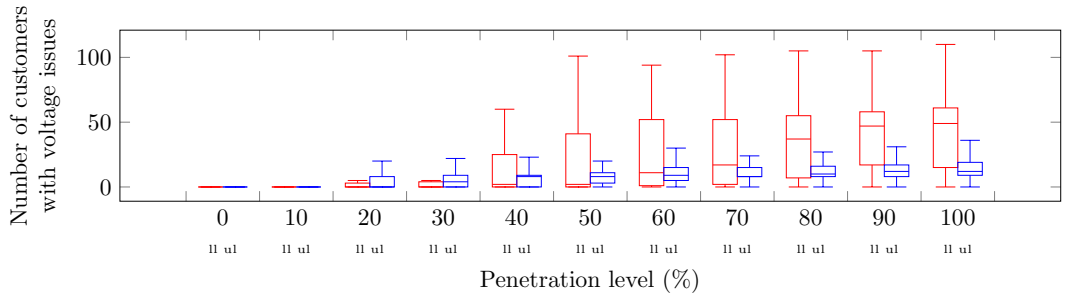
Figure 3.14: Number of customers with voltage issues during weekdays in summer at different penetration levels

the case of generation, a similar approach was taken by scaling the generation profile where voltage showed a higher impact on voltage variation. As an illustration of some of the scenarios investigated and the data obtained, Figure 3.18 shows the time-series data of power injections at one node during the ten-day simulation; Figure 3.18a shows the residential load demand at node 85.C (Scenario S1), and Figure 3.18b shows the power injection from PVs at the same node (Scenario S2). The selection of nodes for scenarios was based on the voltage variation observed in the previous section and the distance from the main feeder (for all scenarios, except for node 48, where the only motivation was the inclusion of three-phase load injection based on the original reference). The nodal voltage and power injection measurements were sampled at a resolution of 10 minutes for all scenarios. This sampling rate was considered realistic for voltage measurements and is illustrated in size between t_k and t_{k+1} in Figure 3.18. In total, 1440 measurements were taken for each variable in each scenario.

The perturbations used in this study were deliberately large in order to maximise their impact on the network, and represent critical scenarios for voltage control. In distribution systems, it is common for several small perturbations to occur simultaneously throughout the network. However, these are often automatically or naturally mitigated and the power across the system remains largely



(a) OLTCs connected and capacitor banks disconnected



(b) OLTCs connected and capacitor banks connected

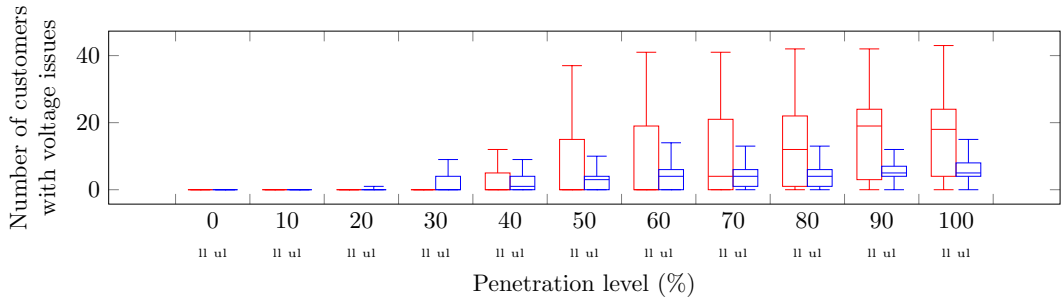
Figure 3.15: Number of customers with voltage issues during weekends in summer at different penetration levels

balanced. The focus of this part of the thesis and the proposed methodology is on scenarios where additional mitigation via control actions is required to avoid voltages becoming excessively high or low. These scenarios also approximate regular operation where one part of the system is showing a high imbalance in power distribution across the nodes.

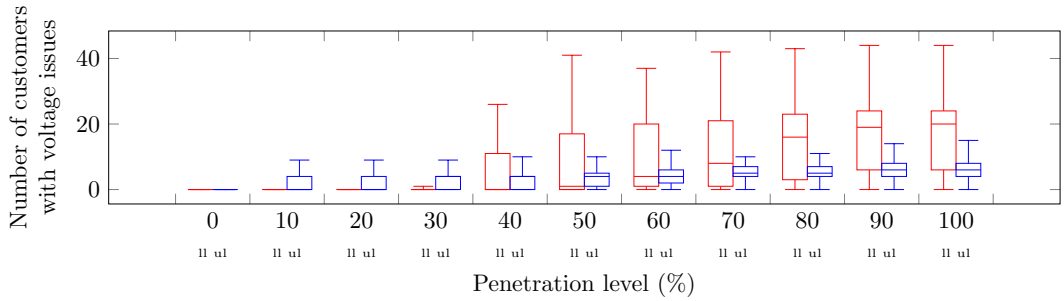
The proposed metric in this section is based on a weighted transformed Pearson correlation coefficient to indicate the most impacted nodes (in terms of voltage variations) from a power injection at a given node. This provides a partial assessment of the *controllability* of the distribution system and potential decision support on which nodal power injections are effective for voltage control.

3.4.2 Preliminary analysis: inferring the voltage–power characteristic from data

A basic characteristic used to describe distribution system behaviour is the X/R ratio. Many existing works on network estimation and identification, e.g. [108, 213], assume a small X/R ratio and therefore that voltage control is achieved by supplying active power to the grid. In practice, however, small X/R may not be a reliable assumption and the supply of reactive power (e.g. by inverters) could



(a) OLTCs connected and capacitor banks disconnected



(b) OLTCs connected and capacitor banks connected

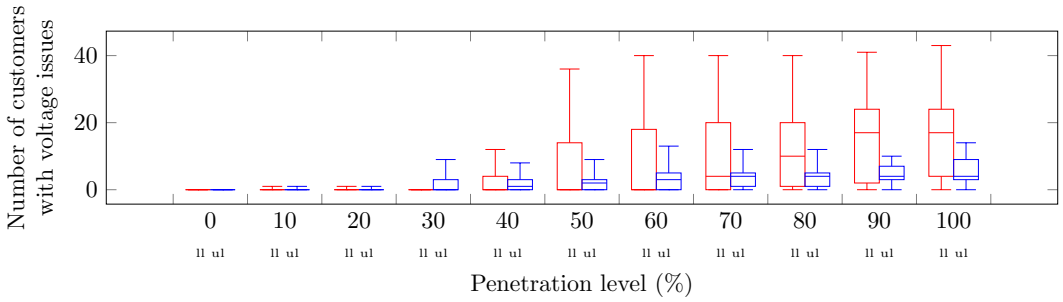
Figure 3.16: Number of customers with voltage issues during weekdays in winter at different penetration levels

be beneficial for voltage control; ultimately this depends on the actual, rather than assumed, voltage–power characteristic of the network. Therefore, in this section a review is given of how to estimate this characteristic from data when impedance parameters are unknown.

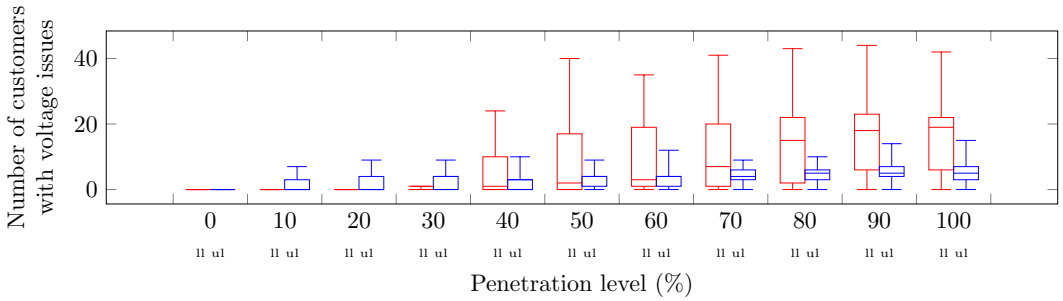
Even with the X/R ratio not known, voltage and power measurements provide information about network properties and how voltage control should be actuated, as the following simple analysis shows. Consider the reduced two-node equivalent system in Figure 2.2. Equation (2.4) can be rewritten in terms of the line loss as

$$\begin{aligned} S^{\text{losses}} = \tilde{V}^{\text{line}} \tilde{I}_r^* &= (R_s + \jmath X_s) \tilde{I}_r \tilde{I}_r^* \\ &= |I_r|^2 R_s + \jmath |I_r|^2 X_s, \end{aligned} \tag{3.3}$$

where $\tilde{V}^{\text{line}} \coloneqq \underline{V_s/\varphi_s} - \underline{V_r/\varphi_r}$ is the voltage drop over the line. Since the same current magnitude determines both real and imaginary parts, it can be concluded that there are proportional relationships between the active component of the impedance and the active power ($P^{\text{losses}} \propto R_s$) and the reactive component of the



(a) OLTCs connected and capacitor banks disconnected



(b) OLTCs connected and capacitor banks connected

Figure 3.17: Number of customers with voltage issues during weekends in winter at different penetration levels

impedance and the reactive power ($Q^{\text{losses}} \propto X_s$). Moreover,

$$P^{\text{losses}} > Q^{\text{losses}} \implies R_s > X_s \tag{3.4}$$

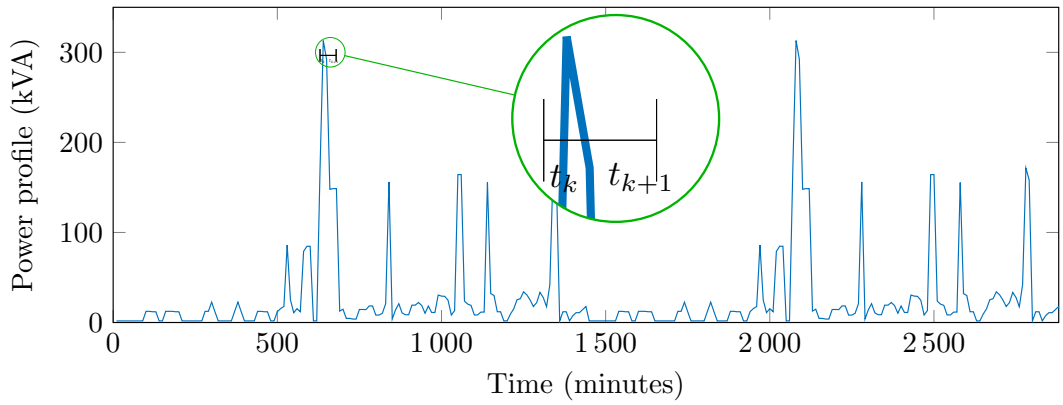
$$P^{\text{losses}} < Q^{\text{losses}} \implies R_s < X_s \tag{3.5}$$

$$P^{\text{losses}} \approx Q^{\text{losses}} \implies R_s \approx X_s \tag{3.6}$$

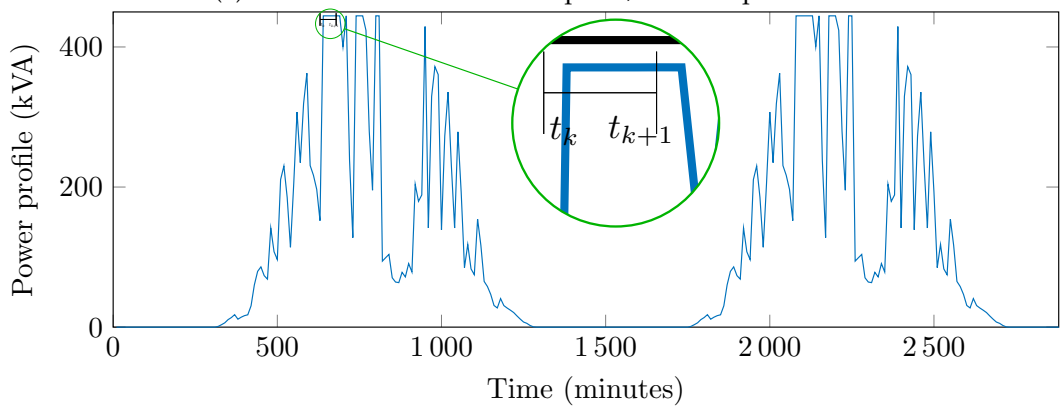
Therefore, determining the power dissipated or stored in lines at discrete points in the network can indicate the effective X/R ratio at those points and therefore what kind of power injection is most effective for local voltage control. On the other hand, such an approach offers no information on which nodal voltages are effected, and to what extent, by nodal power injections; it is required for that a more comprehensive model of the voltage–power relationship.

3.4.3 Validations for power flowing through lines

To query this assumption and attempt to expose the R/X ratio for the test network during the scenarios defined in the previous section, the nodal voltages and line power injections were measured during each of the described perturbations, and correlations between nodal voltage variations and line power dissipations/stor-



(a) Scenario S1: Load consumption, node 85 phase C



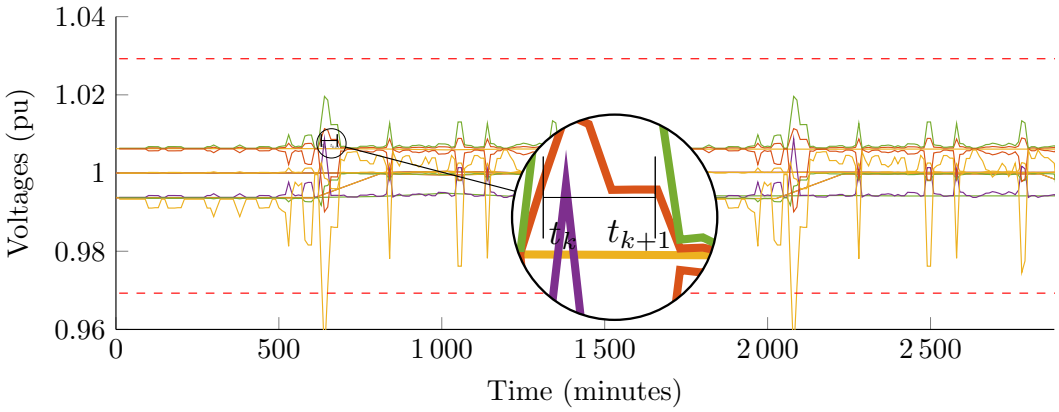
(b) Scenario S2: Photovoltaic generation, node 85 phase C

Figure 3.18: Power profiles at node 85 phase C over ten days, scenarios S1 and S2

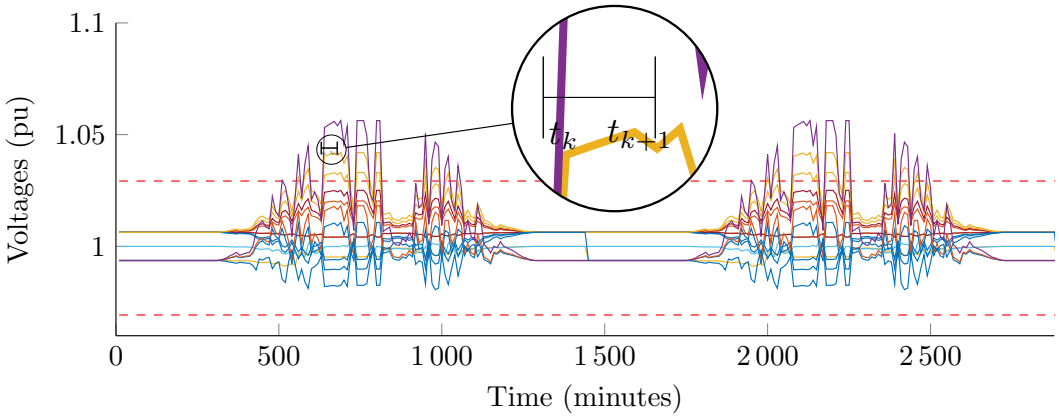
ages were examined.

First, the voltage profiles for scenarios S1 and S2 are presented in Figure 3.19a and Figure 3.19b, respectively. The action–reaction effect of the power–voltage relation is clearly observed in both results when compared with the applied power perturbations shown in Figure 3.18. In particular, it can be seen that in Scenario S2, wherein a significant amount of power is injected by PVs at node 85.C, the network experiences a significant overvoltage event at more that one node; conversely, and as expected, significant undervoltages are observed during Scenario S1.

To analyse the simulation results further, the power and voltage data are time-windowed to extract only those data corresponding to significant voltage events. Let $e_{t_k} \coloneqq [t_k, t_{k+1}]$ denote the time window starting at time $t = t_k$ and extending to time $t = t_{k+1}$. The corresponding windows of interest are indicated in Figures 3.18 and 3.19: in Scenario S1, for example, $t_k = 640$ minutes, while in Scenario S2 $t_k = 690$ minutes. In both cases $t_{k+1} - t_k = 60$ minutes, *i.e.* each window is one-hour long and—considering that the data are sampled every ten minutes—contains



(a) Scenario S1 (Load perturbation)



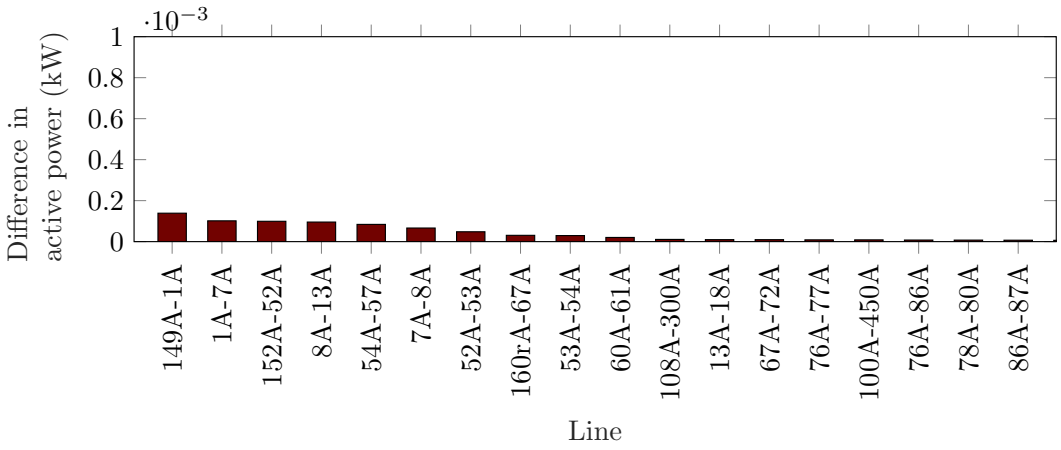
(b) Scenario S2 (PV injection)

Figure 3.19: Voltages at all network nodes during two simulated days of scenarios S1 and S2. The ±3% off-nominal voltage limits are indicated by dashed lines

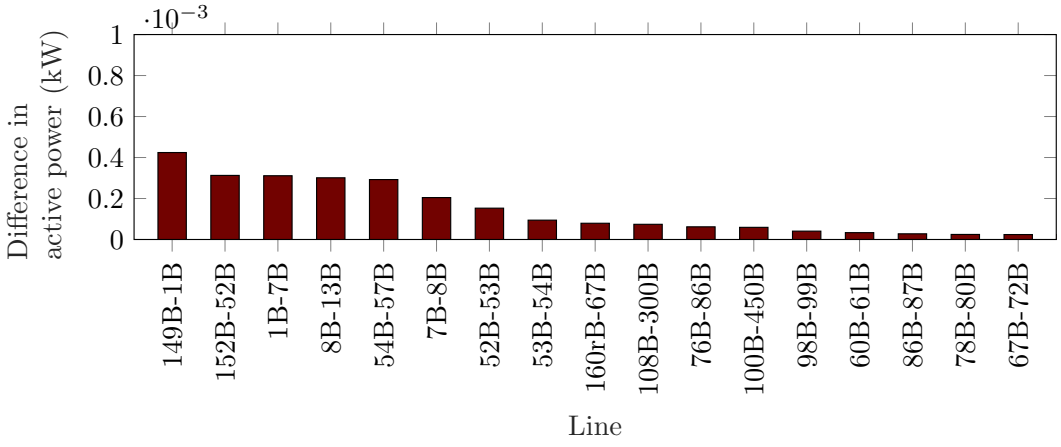
six samples. This choice achieves a reasonable balance between capturing the quasi-dynamics of interest and data storage requirements for the scenarios under consideration.

Figures 3.20 and 3.21 then illustrate, for Scenario S1, the difference in active power dissipation and reactive power storage in lines during the significant undervoltage event that began at t_k = 640 min (i.e., the difference between the maximum and minimum values of dissipated power observed during that event). The lines are sorted in descending order of the magnitude of the observed power dissipation or storage, and only the top 18 lines are presented for each phase.

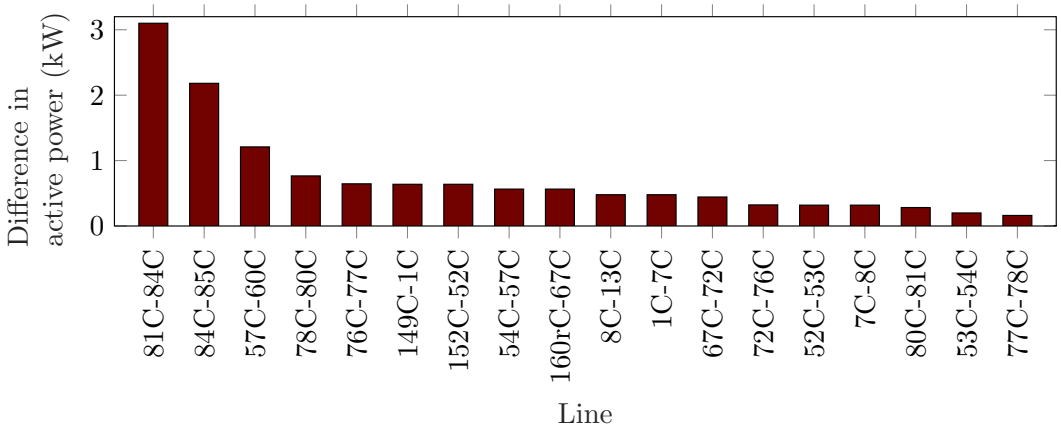
Similarly Figure 3.22 shows, for the same scenario, the 16 nodes that experienced the largest voltage variation on each phase (compared to the nominal value, and calculating the difference between the maximum value and the minimum value experienced during the event e_k). For Scenario S2, similar distributions



(a) Active power through lines on phase A

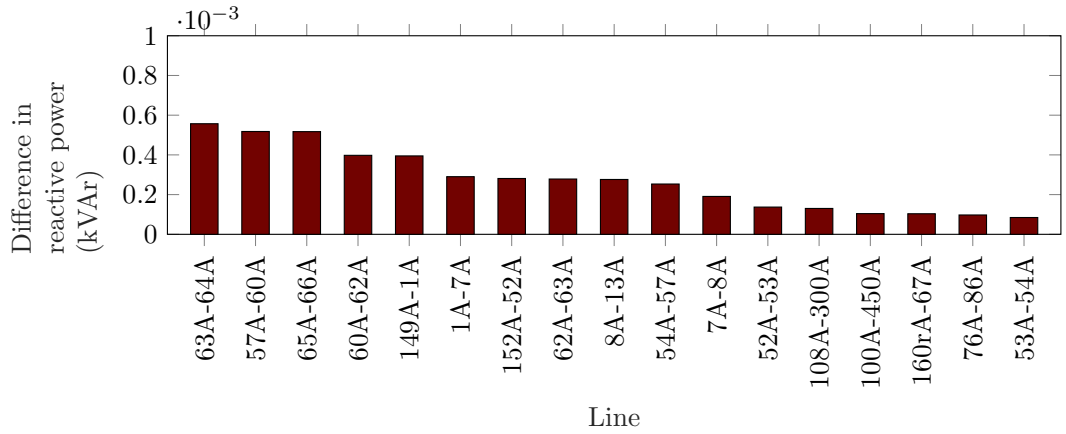


(b) Active power through lines on phase B

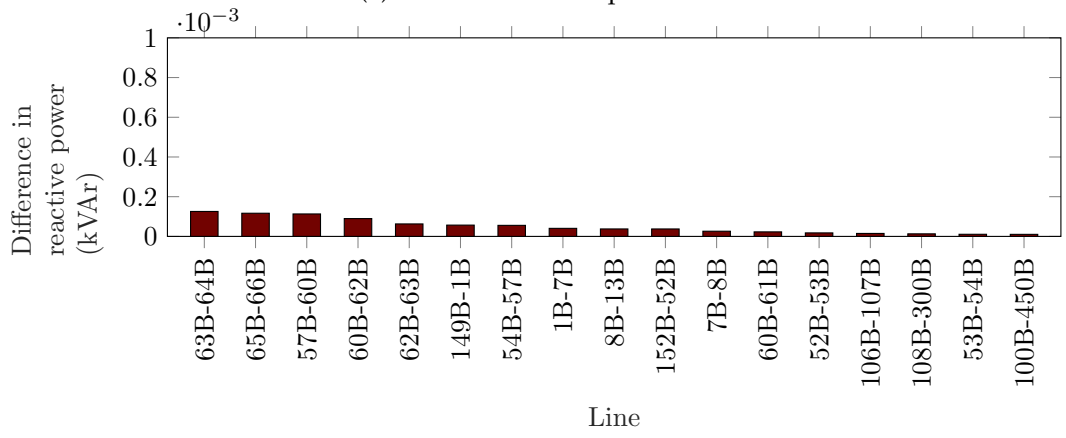


(c) Active power through lines on phase C

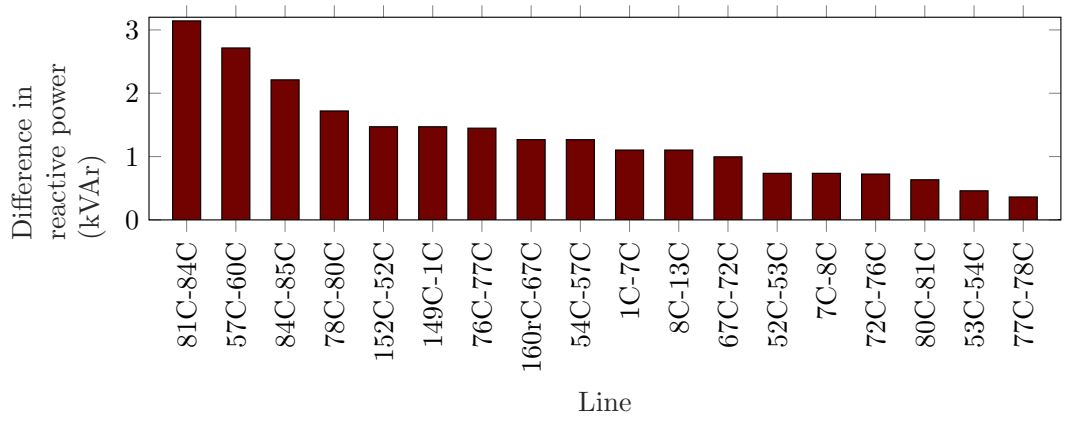
Figure 3.20: Difference in active power flowing through relevant lines of the system during the event when there is a high perturbation at node 85 phase C (Scenario S1, t_k = 640min)



(a) Relevant lines on phase A



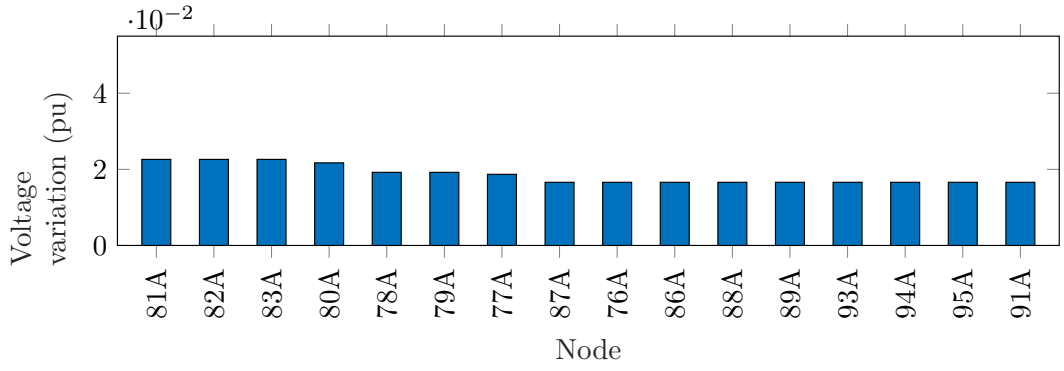
(b) Relevant lines on phase B



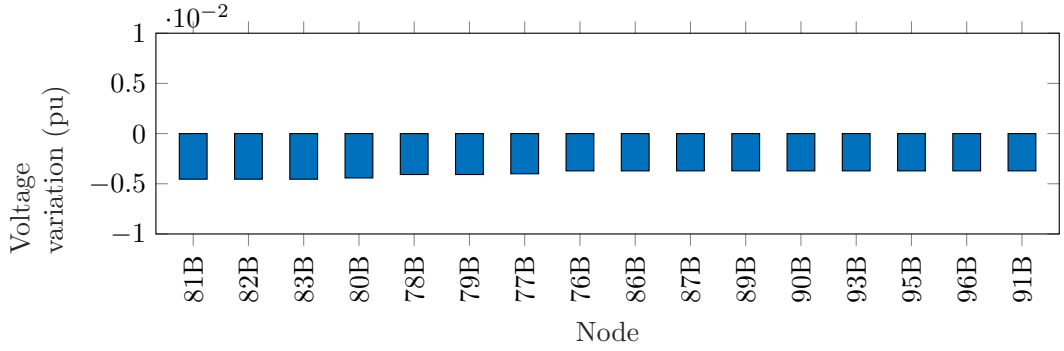
(c) Relevant lines on phase C

Figure 3.21: Difference in reactive power flowing through relevant lines of the system during the event when there is a high perturbation at node 85 phase C (Scenario S1, t_k = 640min)

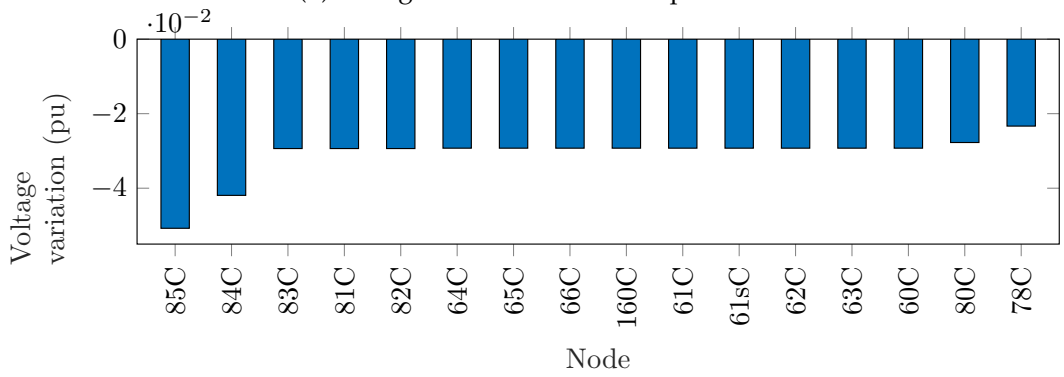
of line power losses/storages were obtained but the voltage variations are positive on the same phase where perturbation is done, which is consistent with PV injections—see Figure 3.23.



(a) Voltage variation results on phase A



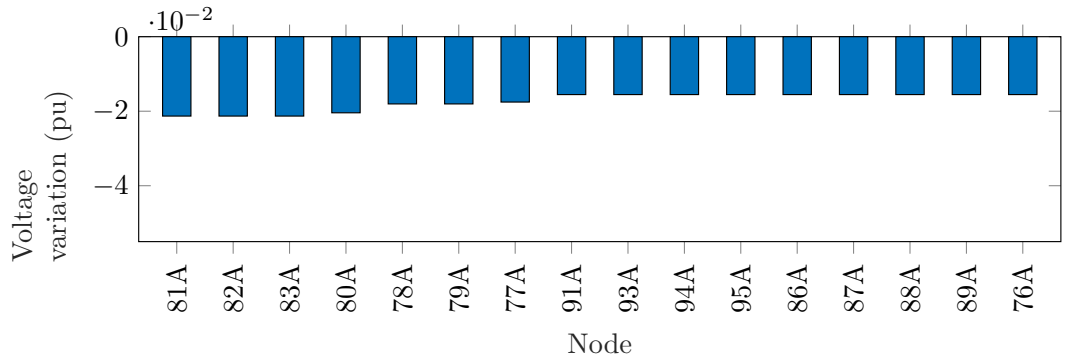
(b) Voltage variation results on phase B



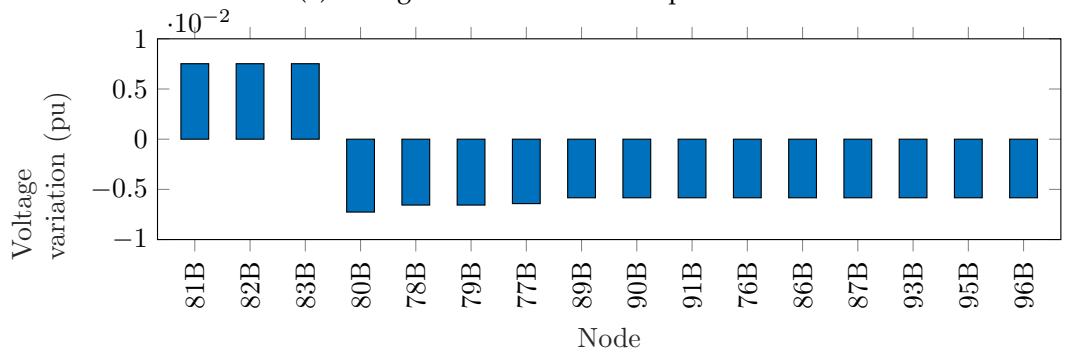
(c) Voltage variation results on phase C

Figure 3.22: Voltage variations at relevant nodes on each phase during the event when there is a high perturbation (consumption) at node 85 phase C (Scenario S1, t_k = 640min)

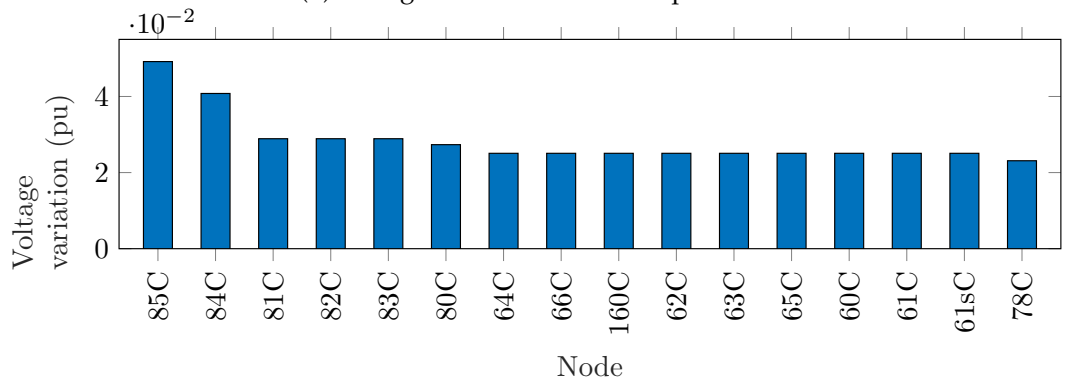
It is shown that the most affected nodes in terms of voltages are also the start or end nodes in the lines that dissipate or store the most power; for example, nodes 85.C, 84.C are the two most impacted nodes and also feature in the most



(a) Voltage variation results on phase A



(b) Voltage variation results on phase B



(c) Voltage variation results on phase C

Figure 3.23: Voltage variations at relevant nodes on each phase during the event when there is a high perturbation (generation) at node 85 phase C (Scenario S2, t_k = 690min)

affected lines. It can also be observed that for each line the active power dissipated and reactive power stored are similar in magnitude; since, as pointed out in equation (3.3), this indicates that for this network under these scenarios the R/X ratio may be around unity.

From this simple example, it can be concluded that a general assumption about the R/X ratio of a distribution network—and whether voltage control would be

most effectively achieved by active or reactive power injections—should not be made without analysing information provided by measurements. On the other hand, the same example shows that measuring power dissipated or stored in lines and correlating with voltage measurements can provide some indication of the R/X ratio even without having system topology information.

3.4.4 Estimating the network power–voltage characteristic from data

System voltages and the power flows lie on the manifold characterised by the nonlinear power flow equations. In practice, however, voltages and flows may be confined to only a small region of the manifold around the operating point, determined by permitted operational limits. This motivates and justifies the use of linearisation to describe voltage–power relationships. In [214], linearisation of the power flow equation for radial distribution systems leads to the voltage–power relation

$$\bar{V} = \mathbf{1} + \bar{\Phi_R}\bar{P} + \bar{\Phi_X}\bar{Q}, \tag{3.7}$$

where $\bar{V}$ is the vector of voltages for nodes under analysis (without loss of generality, it is assumed that the voltage at substation is 1 pu), $\bar{P}$ and $\bar{Q}$ are matrices of active and reactive power injections, and $\bar{\Phi_R}$ and $\bar{\Phi_X}$ are matrices of sensitivities of nodal voltages to active and reactive power injections, respectively; for the voltage at node i considering power injections at node j

$$\Phi_{R_{ij}} = \frac{\partial V_i}{\partial P_j} \quad \text{and} \quad \Phi_{X_{ij}} = \frac{\partial V_i}{\partial Q_j}. \tag{3.8}$$

Where distributed generation is present in the network, equation (3.7) is easily modified to account for those nodes (denoted g) providing power and those that constitute loads (denoted d):

$$\bar{V} = \mathbf{1} + \bar{\Phi_R}\bar{P}^g + \bar{\Phi_X}\bar{Q}^g - \bar{\Phi_R}\bar{P}^d - \bar{\Phi_X}\bar{Q}^d. \tag{3.9}$$

These expressions then provide the desired information about how active and reactive power injections affect nodal voltages throughout the network. However, computing the sensitivity matrices $\bar{\Phi_R}$ and $\bar{\Phi_X}$ relies on either knowing the system topology [215, 216] or, if estimated from measurements, that these sensitivities are time invariant [217–219]. Thus, equations (3.7) and (3.9) represent a traditional load flow model, whereas the voltage–power characteristic in a modern distribution system is governed by the individual and combined quasi-dynamic behaviours of the electrical system, loads and generation; for example, the actions of consumers within and across days, and the daily and diurnal variations in solar

and wind availability. Therefore, an alternative approach is needed if the model should capture these effects.

3.4.5 The Pearson correlation as a tool to identify connectivity

The previous analyses provide methods for characterising the voltage–power relation in a network from measurements, either in the form of the X/R ratio at discrete points or the linear whole-system model (3.9); the latter quantifies how power injections affect nodal voltages, but assumes time invariance of this sensitivity. The aim is to establish a similarly informative relationship, using available measurements of power injections and voltages, albeit also capturing the time-varying effects of loads and DG. To this end, the first step is to identify the nodal connectivity in the network in order to define which nodal voltages are affected by a power injection.

An effective approach to identifying this connectivity is using the statistics available from real-time voltage measurements; for example, previous works have used signatures in time-series data to identify topology changes [220] and the Pearson correlation coefficient as an indicator of phase identification and connectivity [108, 221, 222]. This latter idea is adopted here and extended in the next section to allow determination of the sensitivity of voltages to power injections. First, it was reviewed this technique and illustrate its usefulness for connectivity and phase identification.

Suppose $V_i(t_k, t_{k+1})$ denotes the time series of N_t measurements of the voltage at node i sampled with period T seconds over a window between times t_k and $t_{k+1} = t_k + (N_t - 1)T$, *i.e.* $\{V_i(t_k), V_i(t_k + T), V_i(t_k + 2T), \ldots, V_i(t_{k+1})\}$. The (sample) Pearson coefficient relating nodes i and j is

$$\rho\left(V_i(t_k, t_{k+1}), V_j(t_k, t_{k+1})\right) \coloneqq \frac{\sum_{\ell=0}^{N_t-1} \left(V_i(t_k + \ell T) - \bar{V}_i\right)\left(V_j(t_k + \ell T) - \bar{V}_j\right)}{\sqrt{\sum_{\ell=0}^{N_t-1} \left(V_i(t_k + \ell T) - \bar{V}_i\right)^2}\sqrt{\sum_{\ell=0}^{N_t-1} \left(V_j(t_k + \ell T) - \bar{V}_j\right)^2}} \tag{3.10}$$

where

$$\bar{V}_i \coloneqq \frac{1}{N_t}\sum_{\ell=0}^{N_t-1} V_i(t_k + \ell T) \tag{3.11}$$

is the sample mean of the time series $V_i(t_k, t_{k+1})$ for node i and $\bar{V}_j$ is defined accordingly.

From a geometrical perspective, the Pearson coefficient corresponds to the co-

sine of the angle between the two random variables [223, 224]. Therefore, this coefficient reflects how closely the variations in signals V_i and V_j are matched: if $\rho(V_i, V_j) = 1$ then all of the variance in V_j is explained by V_i. In literature, there are other indices that are often used to develop a sense of the strength and direction of the relationship between two variables, e.g., Spearman coefficient [158, 225] but they are based on different principles. For instance, Spearman coefficient measures monotonic relationships, in which two variables move together in the same direction, and not necessarily at a constant rate. Additionally, it works with ordinal data, which is not the data that is expected to be necessarily used in this context. On the other hand, Pearson coefficients are used to measure linear relationships and is sensitive to small changes, which can be potentially useful in the definition of a new metric.

To give a simple visualisation of how the Pearson coefficients may be used to assess the impact of power injections and identify nodal connectivity, it is calculated these for the IEEE 123-node system, following the pertubations presented on each scenario introduced in previous section. All Pearson coefficients obtained when there is only OLTCs activated are presented in Appendix F. Some of the relevant results are discussed in this section.

Using a node (149.C) close to the feeder as a reference, it was calculated the Pearson coefficient between each node and this reference when there is a single-phase perturbation at node 85.C (Scenario S1). The results are displayed in Figure 3.24. For this and the following examples, unless otherwise specified, the number of measurements per event used to calculate the Pearson correlations using expressions (3.10) and (3.11) corresponds to 60 minutes (7 samples per calculation). Values used for calculations and results of this case are shown in Appendix F, table F.1. This figure shows the distribution system considering a cable for each phase (where applicable). It is observed that in the main, there are cables with three phases, while some phases are not available for some of the downstream parts. The colour of each line represents the correlation obtained with respect to the indicated voltage node and phase (some of them will have a positive correlation, while others will have a negative value, which represents the phase in which the voltage is measured). The applied perturbation and voltage change at 85.C causes highly correlated voltage changes at all other nodes in the network. (Taking any other reference node shows a similar result.) Two of the phases (B and C) respond to the applied perturbation with positive correlation while the other (A) shows an inverse response because of the electromagnetic compensation in the system. From these data, it can be readily identified that the network is fully connected. More than that, however, this brief example shows that measured changes

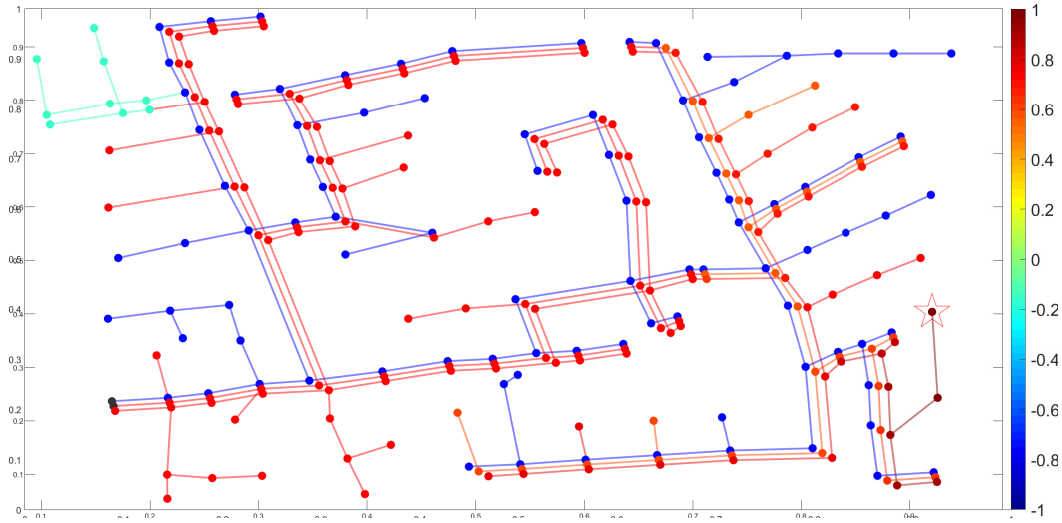


Figure 3.24: Pearson coefficients obtained for each node when there is an event with high perturbation at node 85, phase C

in voltages provide information about the phase in which a perturbation occurred.

Another example to understand the impact of power injections corresponds to perturbations done in different combinations of phases of the distribution system. The response in a 3-phase balanced perturbation at node 48 is shown in Figure 3.25. Values used for calculations and results of this case are shown in Appendix F, table F.5. This perturbation, applied to all three phases, resulted in the same voltage change across all nodes in the system. This is because the measured voltages are highly correlated, with almost the same values from the main feeder to all the others OLTCs. After this point, the correlation value change (for the case after node 160, it drops close to 0) but it remains the same for all nodes and phases. Therefore, knowledge of one phase is sufficient to control the other two, as is the case in a balanced distribution system, which can be modelled as a single-phase system. Identifying the distinct phases will not make a difference if the perturbation or compensation is applied in the same way.

The last case corresponds to the perturbation done in two of three phases. A power-injected perturbation is done only at nodes 82.A and 85.C. The responses using different references are presented in Figure 3.26 and Figure 3.27. Values used for calculations and results of this case are shown in Appendix F, table F.4. The correlations obtained using node 149.C as a reference show that the applied perturbation and voltage cause highly correlated voltage changes at the nodes related to the perturbations. One of the phases (C) responds with high positive correlation values to the applied perturbation, while the other two phases show different behaviours. Phase A reflects voltage variations that cause low correl-

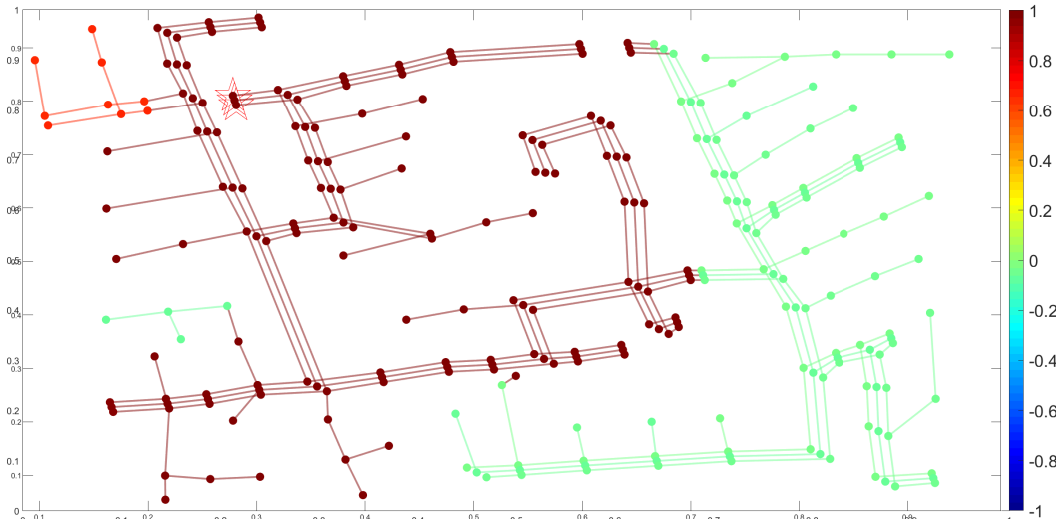


Figure 3.25: Pearson coefficients obtained for each node when there is an event with high perturbation at node 48 (3-phase perturbation)

ation values (close to +0.4), whereas phase B shows negative correlation values (-0.4) in response to the perturbation. Some of the values for the OLTCs that are not close to the perturbation points reflect values that are close to 0 correlation. The results obtained using node 149.A as a reference show something different, where phases A and C both show positive correlation (after the OLTC at node 160) and then low correlation values for the same phases, while phase B shows negative correlation values for the entire network. In both cases, however, it is possible to identify completely the three phases from the distribution system. In this case, each phase can be treated individually, and therefore, each one should be modelled independently.

What is required is to more precisely determine, however, the quantifiable impact on voltages that a perturbation or power injection has. This is addressed in the next section, wherein it is proposed to weight (transformed) Pearson coefficients with measurements of line power flows to produce a metric that indicates the nodes most affected by a perturbation.

3.4.6 A new metric for combined connectivity identification and voltage sensitivity analysis

It was demonstrated, via a simplified analysis in Section 3.4.2, how measurements of active and reactive power flows in lines provide information on the equivalent impedance of the rest of the system. Accordingly, it can be determined whether it is active or reactive power control that would be the more effective means of

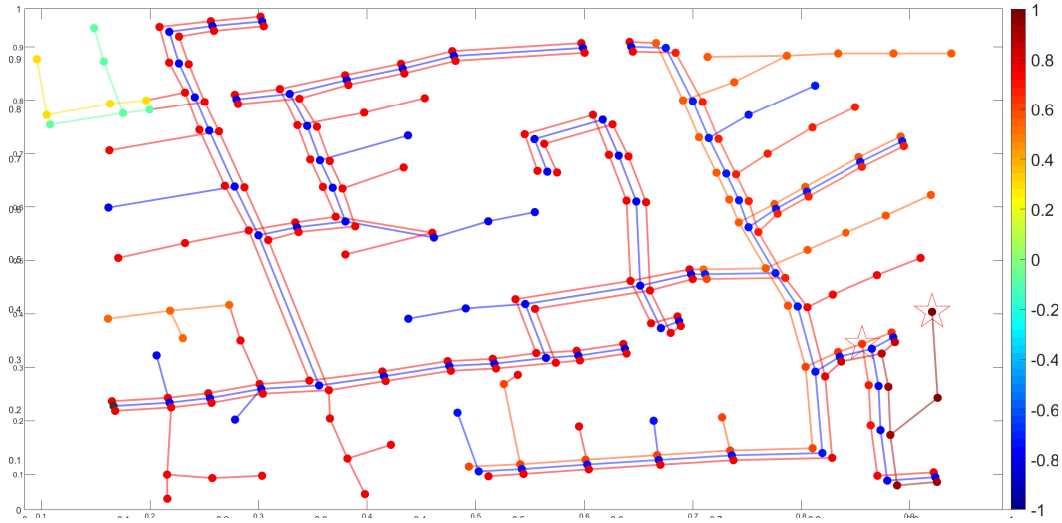


Figure 3.26: Pearson coefficients obtained for each node when there is an event with high perturbation at nodes 82, phase A, and 85, phase C, reference at node 149, phase C

achieving voltage regulation. Moreover, in the previous section, it was demonstrated how the Pearson correlation coefficient enables phase and topology identification from time-series voltage measurements. In this section, these two ideas are combined to produce a single metric that provides information on the connectivity of nodes in the network and indicates the sensitivity of voltages at network nodes to active and reactive power variations at a reference node. The proposed metric thus informs on the extent to which a power injection at a controlled node is able to influence the voltage across the network.

Assume that the voltage is measurable at a certain number of nodes in the network, the set of which is $\mathcal{N}_m \subseteq \mathcal{N}$, and the active and reactive power is measurable at both ends of some set of *paths* $\mathcal{E} \subset \mathcal{N}_m \times \mathcal{N}_m$; a distinction between a *path* and a *line* is presented to accommodate the problem setting of having incomplete topology information, explained as follows.

Figure 3.28 depicts two possible situations for measured nodes n_i and n_k in a radial network. In (a), it is known that n_i and n_j are connected via lines (i,k) and (k,j) and an intermediate node n_k, while in (b) the precise arrangement of lines between n_i and n_j is not known. For (a), the active power "loss" and reactive power "storage" in the path $(i,j) \in \mathcal{E}$ are given, respectively, as

$$P_{ij}^{\text{path}} = P_{ij} + P_{ji} = P_{ij} + P_{ki} + P_{kj} + P_{jk}, \tag{3.12}$$

$$Q_{ij}^{\text{path}} = Q_{ij} + Q_{ji} = Q_{ij} + Q_{ki} + Q_{kj} + Q_{jk}, \tag{3.13}$$

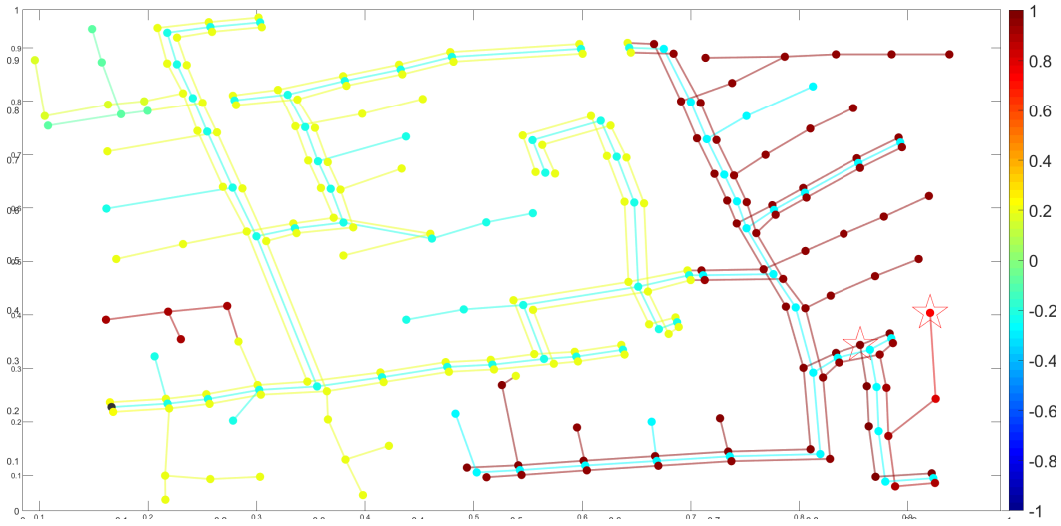


Figure 3.27: Pearson coefficients obtained for each node when there is an event with high perturbation at nodes 82, phase A, and 85, phase C, reference at node 149, phase A

which accounts for the load consumption at k. For (b), the active and reactive path powers are given as

$$P_{ij}^{\text{path}} = P_{ij} + P_{ji}, \tag{3.14}$$

$$Q_{ij}^{\text{path}} = Q_{ij} + Q_{ji}. \tag{3.15}$$

Here P_{ij} is the power injected into the line at n_i in the *direction* of n_j, and *vice versa*; therefore, it is assumed that for $n_i \in \mathcal{N}_m$ the direction of power flows with respect to other measured nodes is *known*, which in turn introduces a tacit assumption on knowledge of the topology of the system but that is not quite as strong as knowing the entire topology. For example, consider the system shown in Figure 3.1 and the problem of calculating P_{ie}^{path} between nodes n_i and n_e; it is required to know whether P_{ec} or P_{er} is the relevant line power injection in the sum. Likewise, to compute P_{ek}^{path} requires knowing whether the direction from n_e to n_k is along (e, c) or (e, r). Therefore, preliminary topology identification from data may be required [108, 109].

The value $S_{ij}^{\text{path}} = P_{ij}^{\text{path}} + \jmath Q_{ij}^{\text{path}}$ accounts for any consumption or generation along the path from n_i to n_j. Even when n_i and n_j are not connected by a line this value provides relevant information on the network with respect to control. For example, assuming n_i and n_j are connected by a path, if the value of S_{ij}^{path} small, then the section in between is near balance, further suggesting that additional control actions are not necessary along that path. On the other hand, if the value is high, this suggests control action (power injections or extractions) could

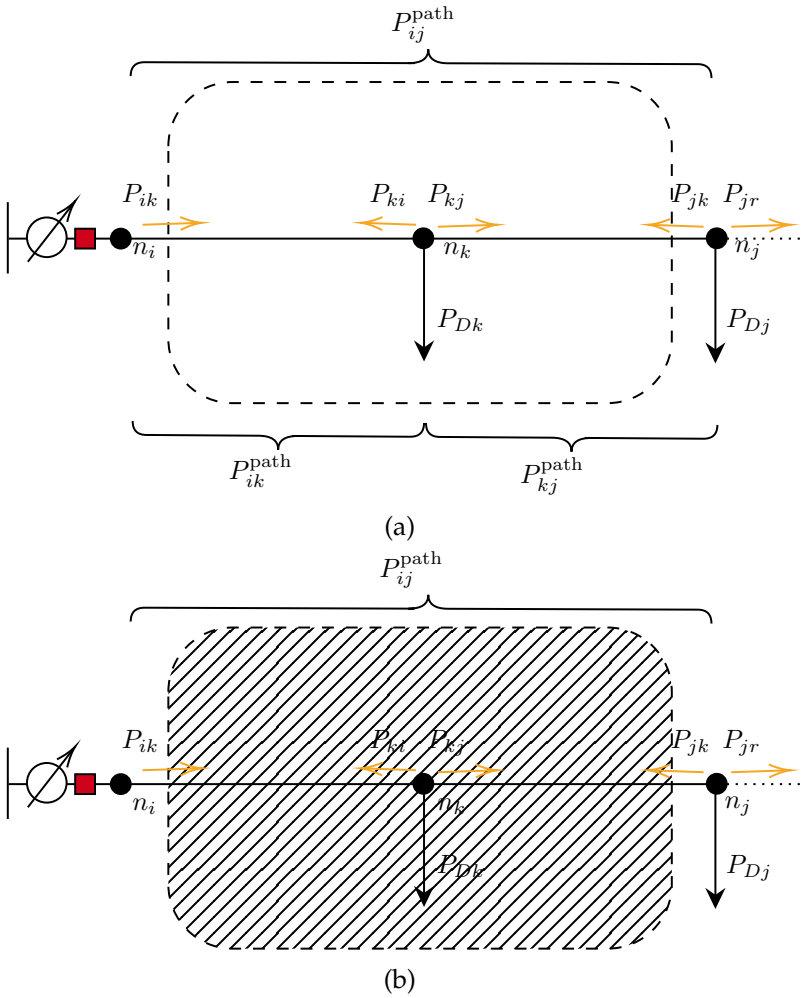


Figure 3.28: Power flows into lines in the cases that the connectivity is (a) known and (b) unknown

be beneficial.

The method applied is highly adaptable, even when there are controlled devices between measurable nodes that cannot be directly measured or controlled. Figures 3.29 and 3.30 illustrate examples of how power balance is considered in the presence of these devices. The operation of the OLTC impacts the balance in $P^{path}ik$ due to associated losses. However, the overall balance in $P^{path}ij$ remains nearly the same even when the connectivity is unknown. Therefore, the value of P_{ij} can be calculated in the same way as indicated in 3.14. Similarly, in the presence of a bank of capacitors, the balance in $Q^{path}ik$ is altered due to the presence of the device operation. However, the overall balance in $Q^{path}ij$ remains almost the same, even when the connectivity is unknown. Therefore, the value of Q_{ij} can be calculated as indicated in 3.15.

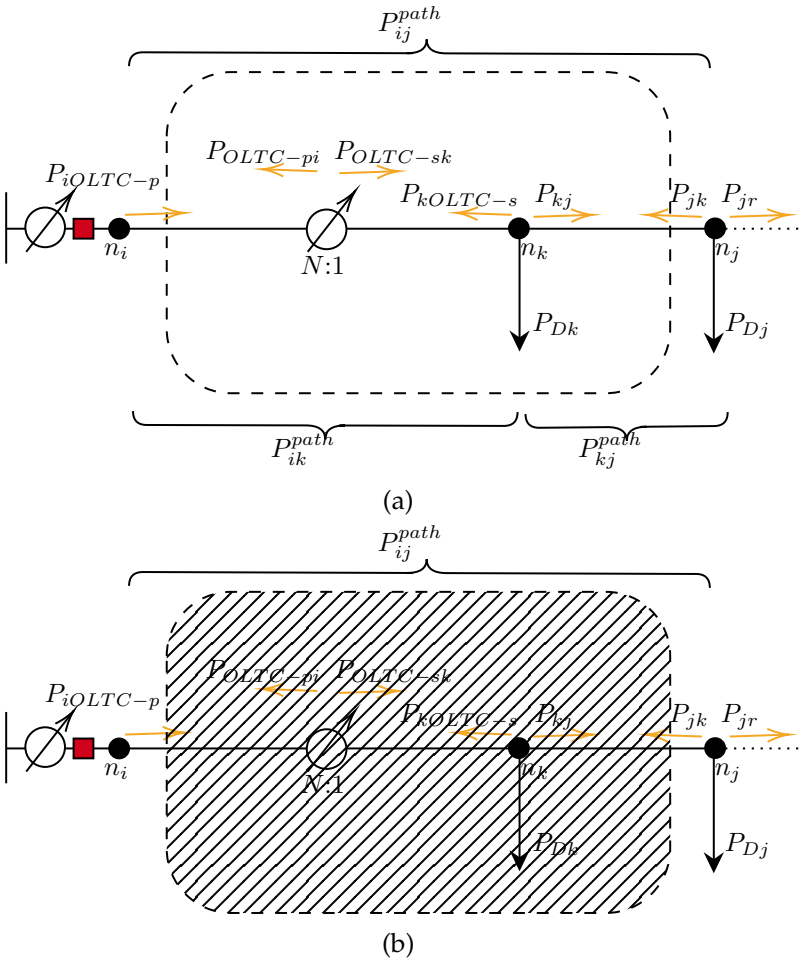


Figure 3.29: Power flows into lines in the cases that the connectivity is (a) known and (b) unknown, considering an OLTC device between measurable nodes

It is proposed to enhance the information on network connectivity and phase information provided by the Pearson correlation coefficients with information provided by these power balances along paths. In particular, it is weighted the Pearson coefficient for (i, j) by the maximum *difference* in P_{ij}^{path} and Q_{ij}^{path} over a time window of observations:

$$\Delta P_{ij}^{\text{path}}(t_k, t_{k+1}) \coloneqq \max_{t\in[t_k,t_{k+1}]}[P_{ij}^{\text{path}}(t_k, t_{k+1})] - \min_{t\in[t_k,t_{k+1}]}[P_{ij}^{\text{path}}(t_k, t_{k+1})], \tag{3.16}$$

$$\Delta Q_{ij}^{\text{path}}(t_k, t_{k+1}) \coloneqq \max_{t\in[t_k,t_{k+1}]}[Q_{ij}^{\text{path}}(t_k, t_{k+1})] - \min_{t\in[t_k,t_{k+1}]}[Q_{ij}^{\text{path}}(t_k, t_{k+1})]. \tag{3.17}$$

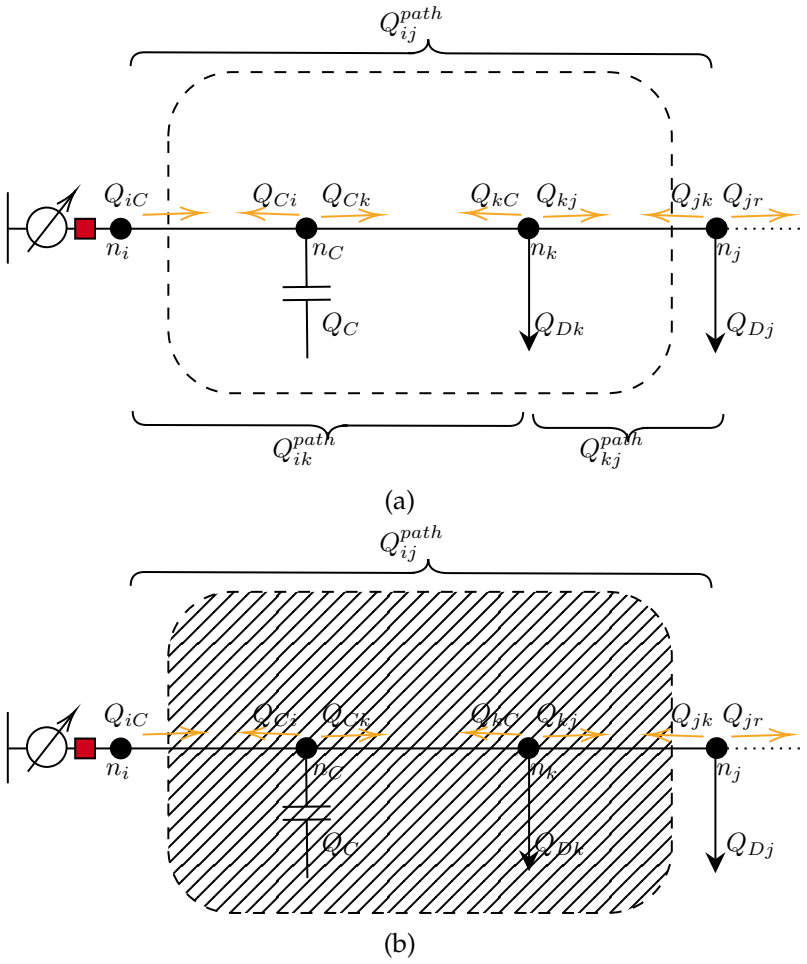


Figure 3.30: Power flows into lines in the cases that the connectivity is (a) known and (b) unknown, considering a capacitor bank between measurable nodes

The (i, j) element of the proposed matrices M^P and M^Q is then defined as

$$M_{ij}^{P}(t_k, t_{k+1}) \coloneqq \Delta P_{ij}^{\text{path}} \mathcal{Z}\left(\rho\left(V_i(t_k, t_{k+1}), V_j(t_k, t_{k+1})\right)\right) \tag{3.18}$$

$$M_{ij}^{Q}(t_k, t_{k+1}) \coloneqq \Delta Q_{ij}^{\text{path}} \mathcal{Z}\left(\rho\left(V_i(t_k, t_{k+1}), V_j(t_k, t_{k+1})\right)\right) \tag{3.19}$$

where the function $\mathcal{Z}(\cdot)$ is the Fisher z-transformation [222]

$$\mathcal{Z}(\rho) = \operatorname{arctanh}(\rho) \tag{3.20}$$

which, when applied to the Pearson correlation value, recovers an approximately normal distribution of coefficients, enhances the values that show higher correlations and can be used for hypothesis testing and confidence interval estimation. In this case, this helps to more clearly distinguish those nodes that are highly

correlated with the reference or perturbation point.

The value of M_{ij}^{P} (M_{ij}^{Q}) is largest when the voltages V_i and V_j are highly correlated (indicating phase and topological connectivity between n_i and n_j) and, simultaneously, the variation in net active (reactive) power between n_i and n_j is large; this is indicative that the large power fluctuations in this path are causing the voltage variations, and control of power injections locally could positively impact voltage. The value is, on the other hand, small if either the voltages are uncorrelated or the power variations are small, in which case control of power injections locally would have minimal effect. (If, for example, path (i, j) has a large value of P_{ij}^{path} but small $\Delta P_{ij}^{\text{path}}$, then it is indicated that voltage variations between n_i and n_j are not caused by the presence of this net power and the voltage variation, if present, is explained by some other effect; therefore, manipulation of power injections locally may have minimal effect on voltages.)

When the connection between two nodes is known and it is done through a line, it is defined that $S_{ij}^{\text{path}} = S_{ij}^{\text{losses}}$. When this path takes a portion of the system that is unknown or non-measured, the value S_{ij}^{path} will represent the power balance between both nodes, which might be or not connect through a line. This provides relevant information of the network.

To illustrate this idea and potential usefulness of these metrics, it is revisited the simple 123-node example with a perturbation at node 85.C. Results obtained for all scenarios are presented in more detail in Appendix F. Some of the relevant results are discussed in this section. Figure 3.31 indicates the resulting paths in Scenario S1 with a values of M^P and M^Q higher than 0.5. This path connects the nodes in which the post-perturbation voltage variation is higher than 0.03 pu. The figure indicates how the effect of the power injection propagates through the distribution system in the form of variations to voltage.

The highest values of M^P and M^Q relates the nodes and lines that are highly correlated to the perturbation observed in measurements. Some of the highlighted paths represented the voltage regulators (connection $67C - 160rC$). This gives a numerical indicator which represents the power that flows between nodes beyond the unperturbed condition (perfect power consumption and power generation balance). This tells about the sections in which the model should pay more attention in order to describe the system under study.

Figure 3.32 indicates the node pairs with the largest values of M^P and M^Q observed for Scenario S1 during the 60-minute undervoltage event starting at t_k = 640 minutes. Firstly, note that Figure 3.32 is a version of Figures 3.20 and 3.21 enhanced with the connectivity and correlation information offered by the transformed Pearson coefficients: Figure 3.32 performs a reweighting and reordering

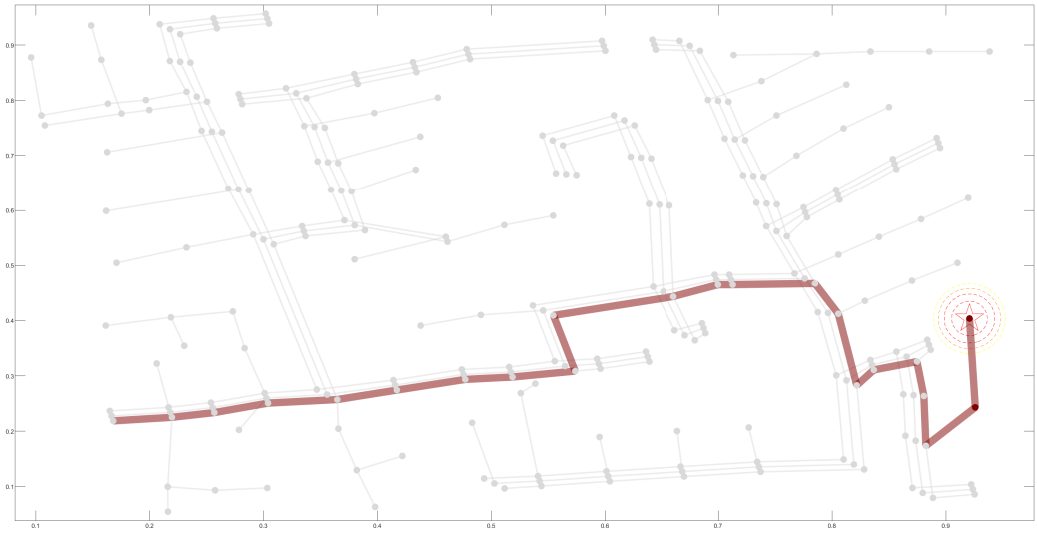


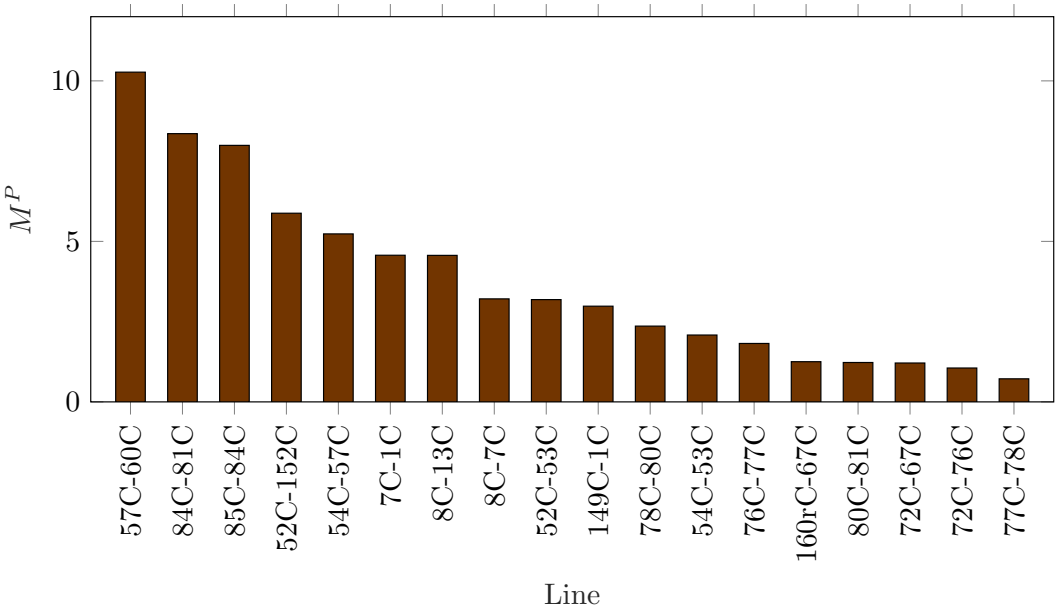
Figure 3.31: Representation of M^P and M^Q values obtained for values higher than 0.5 when there is a high perturbation at node 85.C (Scenario S1)

of the lines that experienced significant power dissipation in Figure 3.20 or storage in Figure 3.21, ranking higher those that have highly correlated end voltages. The most notable example of this is line 149–1, close to the feeder, which sees significant power activity but moves to a lower rank when voltage correlations are taken into account; indeed, neither node 1C nor node 149C appear in the list of nodes that experience significant voltage variations during the perturbation event (Figure 3.22).

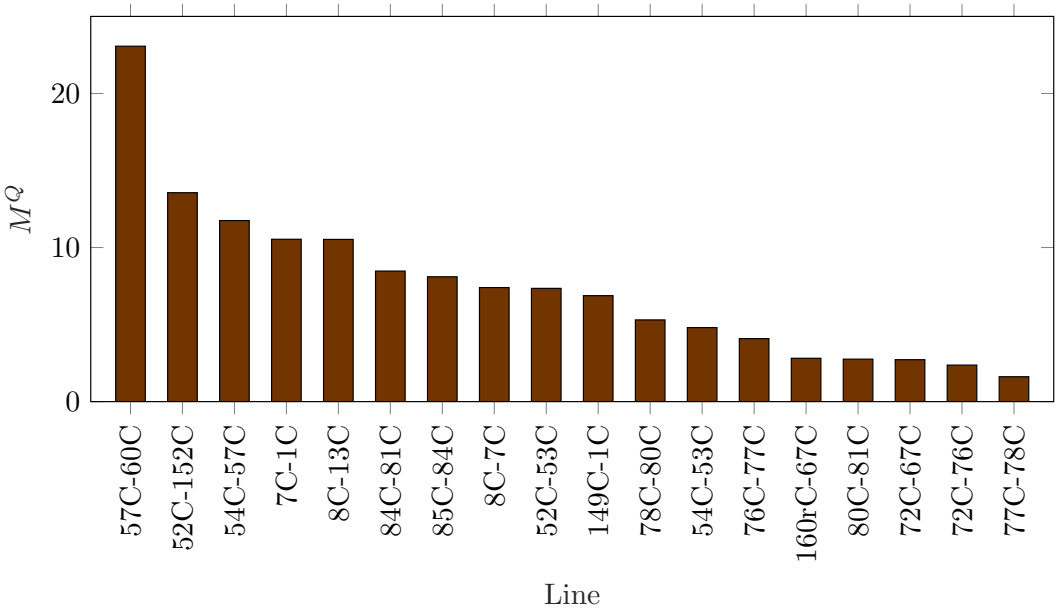
Considering that Scenario S1 corresponds to a load consumption at node 85, phase C, it can be seen that four of the top six lines in terms of magnitude of M^P or M^Q—lines 84–81, 85–84, 80–78, 77–76—are graphically close to the perturbation node and, in particular, are in the radial branch emanating from node 76. All involve nodes that experience significant voltage perturbations (Figure 3.22). The other notable observation is the group of lines along a path that contains voltage regulator at node 160; lines 60–57, 52–152, 67–160, 54–57, 67–72 all rank highly in terms of M^P and M^Q and indicate an area of the network that could benefit from power injection as a control action in response to the large consumption event at node 85.

Figure 3.33 shows the corresponding M^P and M^Q values for Scenario S2, wherein power is injected to node 85C. Similar observations may be made, albeit in the opposite direction: the high M^P and M^Q values are found in lines along paths that link nodes with significant overvoltages (Figure 3.23) and indicate areas of the network where power extractions could be beneficial control actions.

To assess the impact of unmeasured control devices, the same calculation was



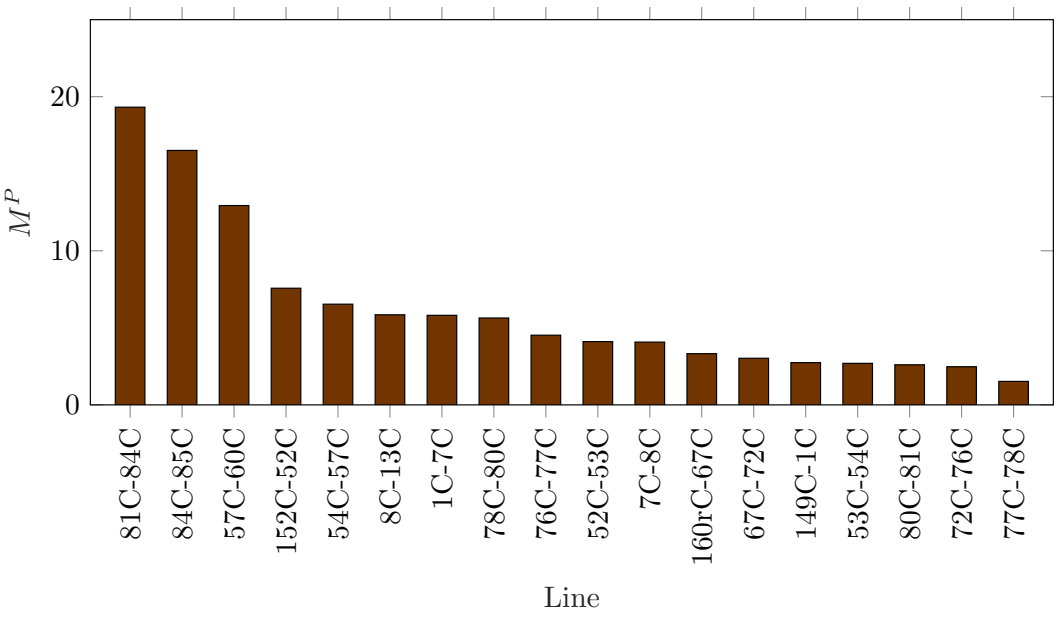
(a) Obtained M^P values



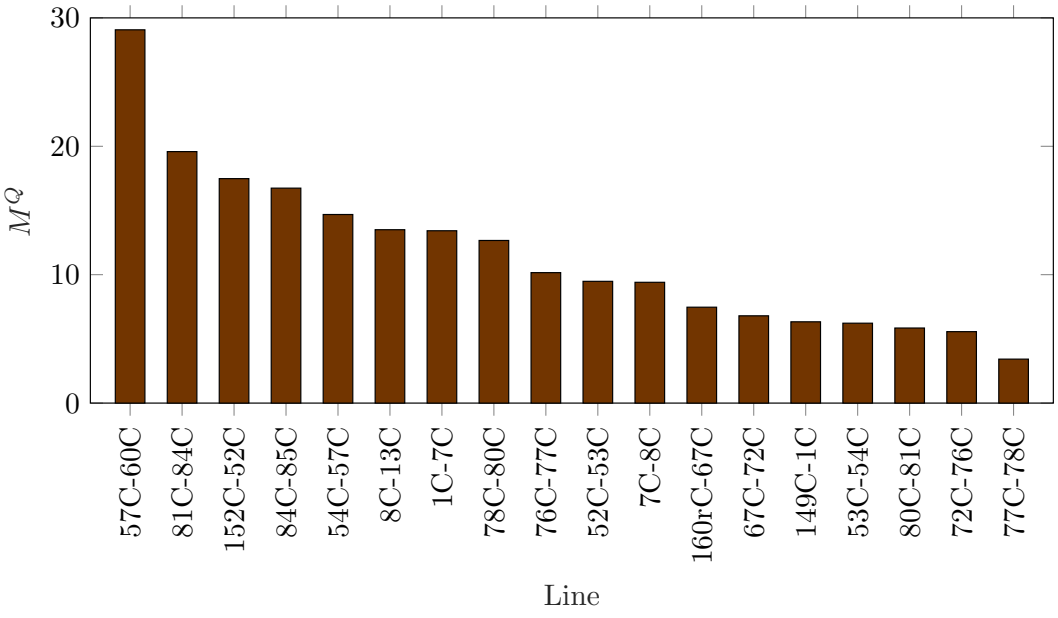
(b) Obtained M^Q values

Figure 3.32: Obtained relevant values M^P and M^Q for Scenario S1 (t_k = 640min)

performed by considering the connection and operation of the capacitor banks installed in the circuit. Figure 3.34 illustrates the resulting paths in Scenario S1 with values of M^P and M^Q higher than 0.5. This time, it is observed that additional nodes in the system are experiencing voltage issues (attributed to the reactive compensation performed in all three phases). Therefore, the metrics M^P and M^Q



(a) Obtained M^P values



(b) Obtained M^Q values

Figure 3.33: Obtained relevant values M^P and M^Q for Scenario S2 (t_k = 690min)

reflect the effects on all phases, in contrast to the previous example where only one phase was affected.

Figure 3.35 indicates the node pairs with the largest values of M^P and M^Q observed for Scenario S1 during the 60-minute undervoltage event starting at t_k = 640 minutes. Figure 3.35 performs a reweighting and reordering of the lines that

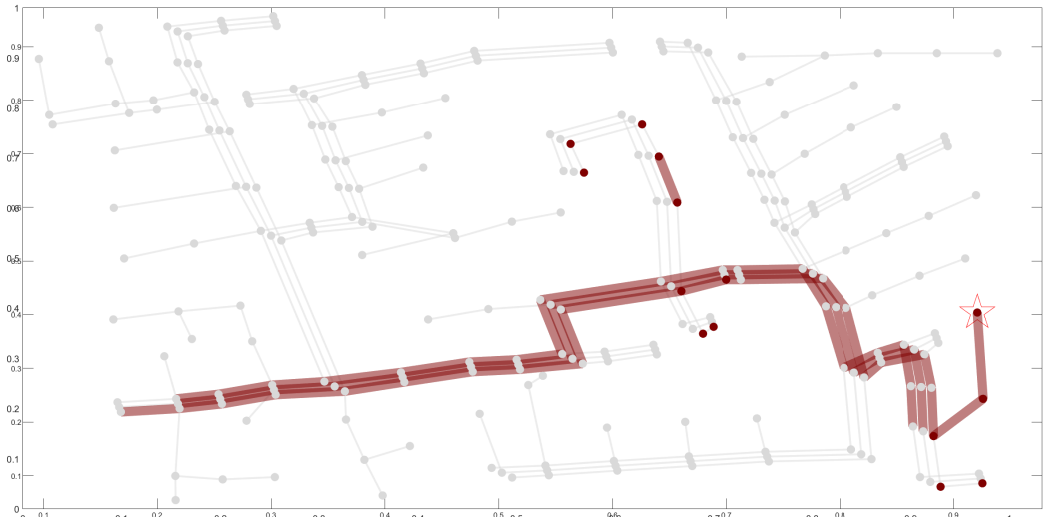
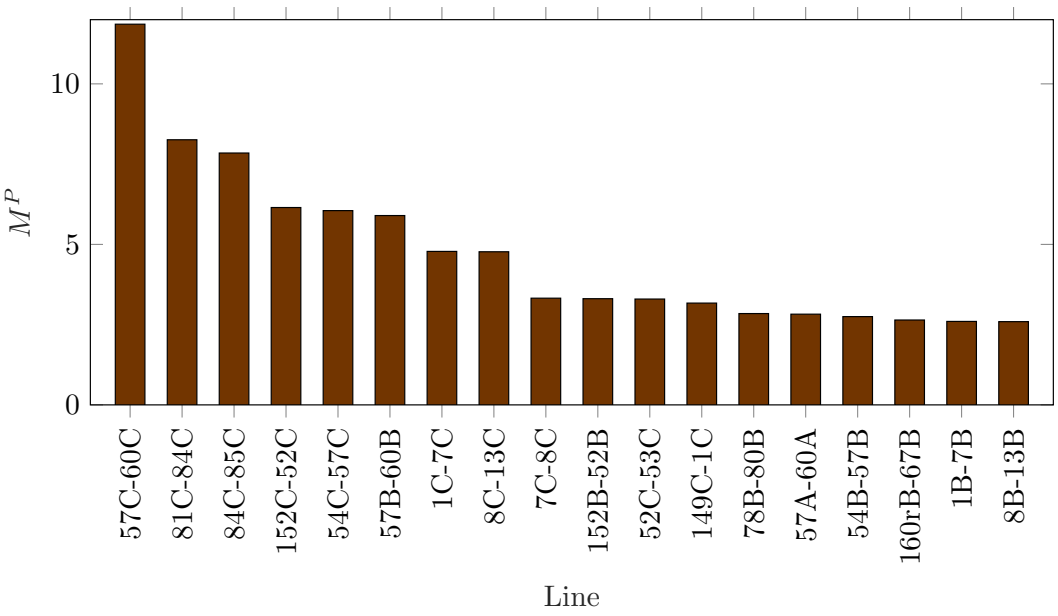


Figure 3.34: Representation of M^P and M^Q values obtained for values higher than 0.5 when there is a high perturbation at node 85.C and the capacitor banks are connected (Scenario S1)

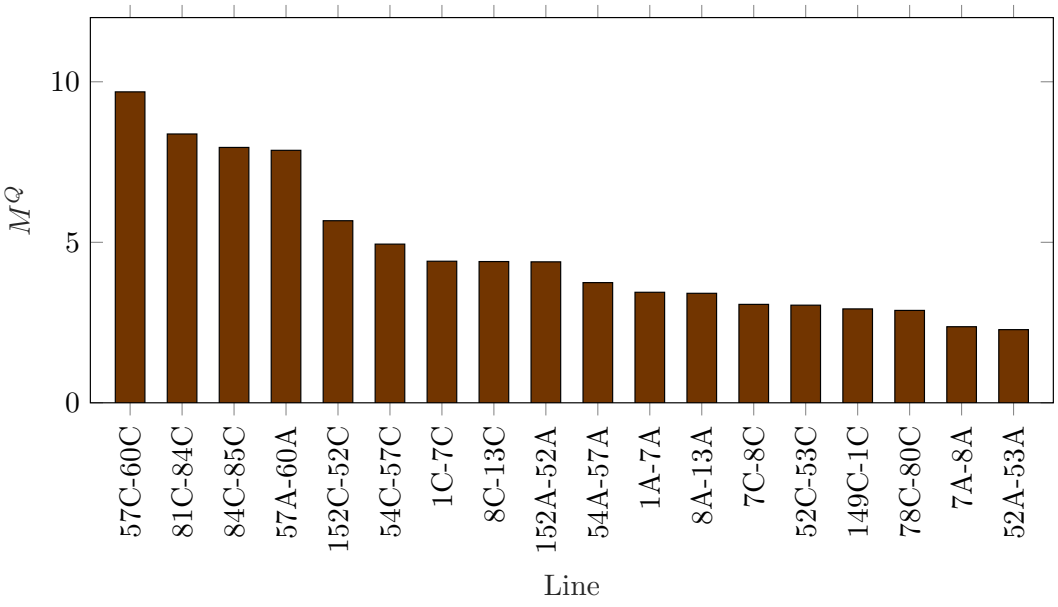
experienced significant power dissipation, ranking higher those that have highly correlated end voltages.

Considering that Scenario S1 corresponds to a load consumption at node 85, phase C, it is observed that four out of the top six lines in terms of the magnitude of M^P or M^Q—lines 84–81, 85–84, 80–78, 77–76—are located close to the perturbation node, specifically in the radial branch originating from node 76. The main difference compared to the previous example is that the metrics now take into account relevant values in other phases, not just phase C. This is due to the impact of reactive compensation occurring in all phases, which significantly affects the voltage not only in phase C. However, this effect is only evident up to node 83, which is the connection point between phase C of node 85 and the rest of the system. Therefore, any compensation performed along the highlighted path will influence the voltage observed in the nodes marked in the figure. Similarly, in line with the previous example, the group of lines along a path that includes the voltage regulator at node 160—lines 60–57, 52–152, 67–160, 54–57, 67–72—ranks high in terms of both M^P and M^Q, indicating an area of the network that could benefit from power injection as a control action in response to the significant consumption event at node 85.

Figure 3.36 shows the corresponding M^P and M^Q values for Scenario S2, wherein power is injected to node 85C. Similar observations may be made, albeit in the opposite direction: the high M^P and M^Q values are found in lines along paths that link nodes with significant overvoltages and indicate areas of the



(a) Obtained M^P values

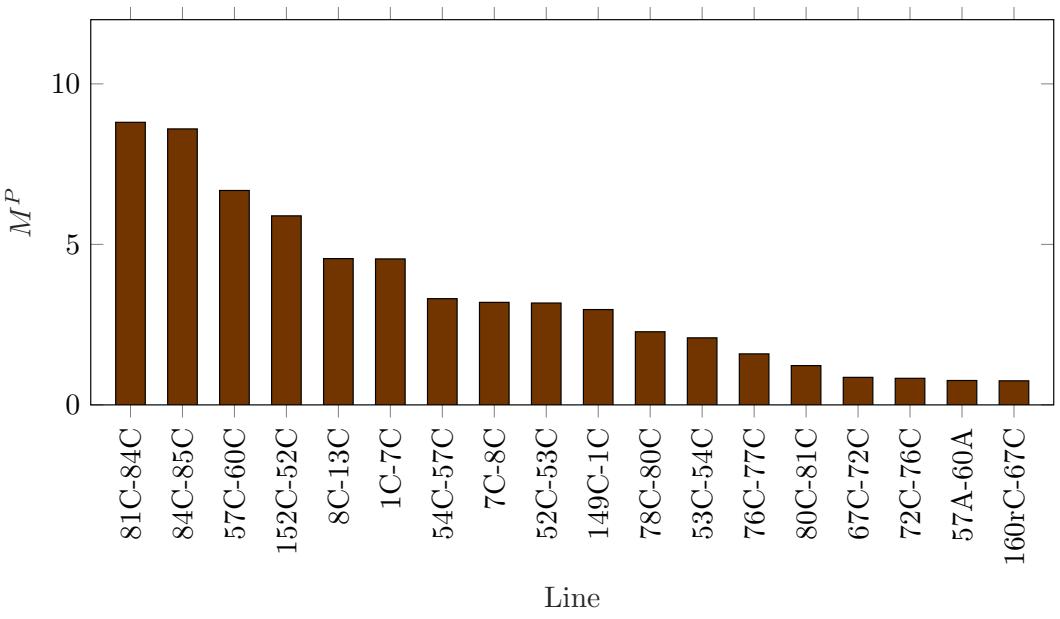


(b) Obtained M^Q values

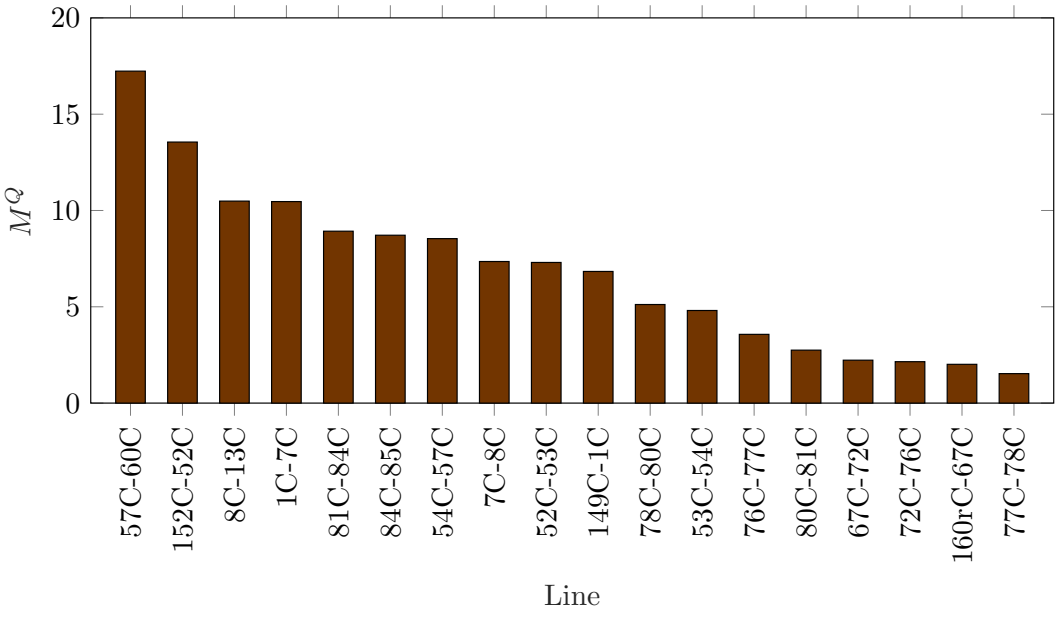
Figure 3.35: Obtained relevant values M^P and M^Q when the capacitor banks are connected for Scenario S1 (t_k = 640min)

network where power extractions could be beneficial control actions.

Additionally, all cases were run with a connection between nodes 54 and 94 was done to evaluate the impact of meshing the circuit. Figure 3.37 illustrates the resulting paths in Scenario S1 with values of M^P and M^Q higher than 0.5. It is shown how some of the lines of phase A that are also part of the meshed region



(a) Obtained M^P values



(b) Obtained M^Q values

Figure 3.36: Obtained relevant values M^P and M^Q when the capacitor banks are connected for Scenario S2 (t_k = 690min)

are highlighted by the calculation. Therefore, the metrics M^P and M^Q reflect the effects on phases A and C, in contrast to the previous cases where only one phase was affected (with only the OLTC connected) or all phases (which included the reactive compensation from capacitor banks in all phases).

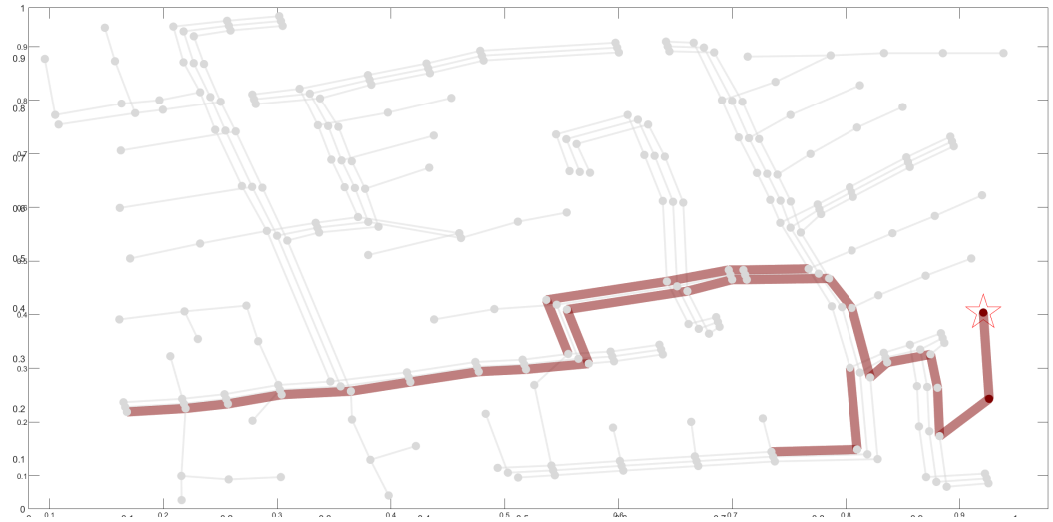
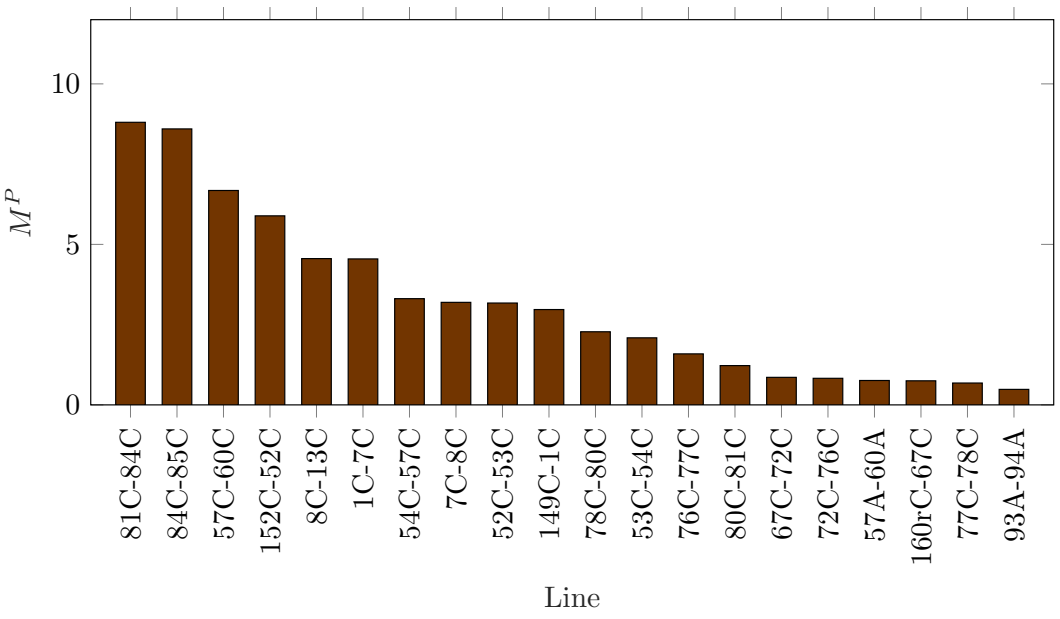


Figure 3.37: Representation of M^P and M^Q values obtained for values higher than 0.5 when there is a high perturbation at node 85.C and meshed by connection between nodes 54 and 94(Scenario S1)

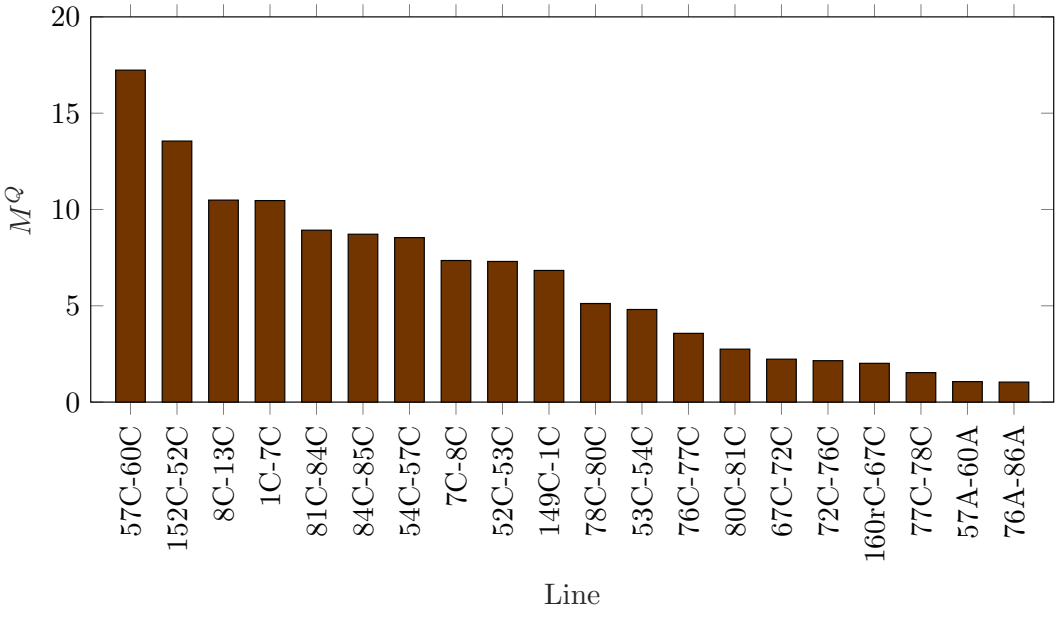
Figure 3.38 indicates the node pairs with the largest values of M^P and M^Q observed for Scenario S1 during the 60-minute undervoltage event starting at t_k = 640 minutes. Figure 3.38 performs a re weighting and reordering of the lines that experienced significant power dissipation, ranking higher those that have highly correlated end voltages.

Considering that Scenario S1 corresponds to a load consumption at node 85, phase C, it is observed, as in previous cases, that four out of the top six lines in terms of the magnitude of M^P or M^Q—lines 84–81, 85–84, 80–78, 77–76—are located close to the perturbation node, specifically in the radial branch originating from node 76. The notable difference is the appearance of lines associated with phase A that connect the node from the event to the point used to mesh the circuit. This indicates that the metrics are capable of capturing the operating condition of the system, whether it is radial or meshed, which is reflected in voltage correlation and power distribution across the circuit. Once again, any compensation performed along the highlighted path will influence the voltage observed in the nodes marked in the figure. Similarly, the group of lines along a path that includes the voltage regulator at node 160—lines 60–57, 52–152, 67–160, 54–57, 67–72—ranks high in terms of both M^P and M^Q, indicating an area of the network that could benefit from power injection as a control action in response to the significant consumption event at node 85.

Figure 3.39 shows the corresponding M^P and M^Q values for Scenario S2, wherein power is injected to node 85C. Similar observations may be made, al-



(a) Obtained M^P values

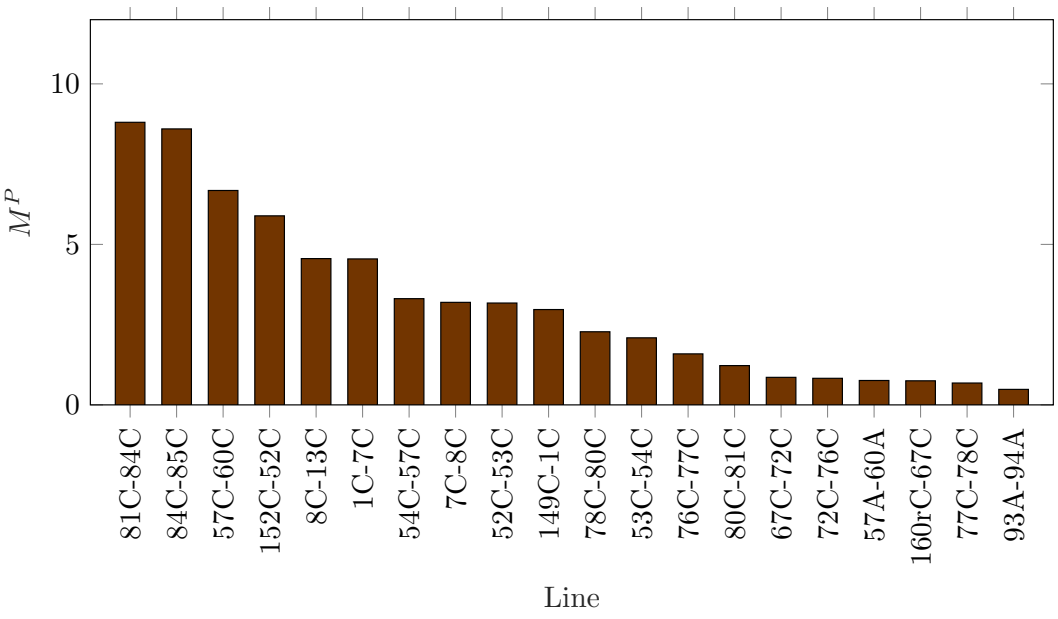


(b) Obtained M^Q values

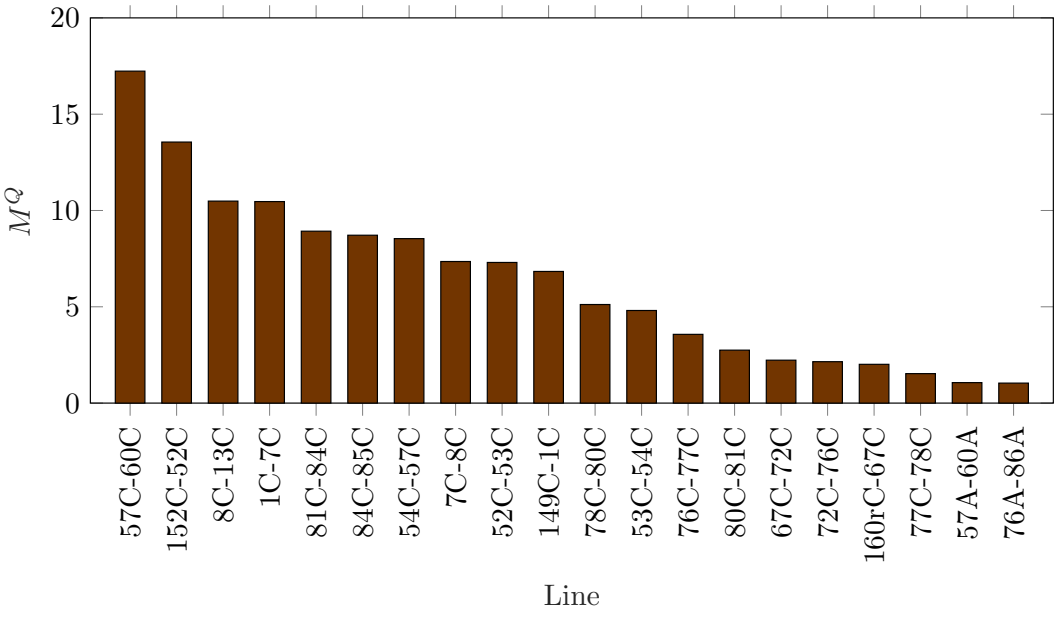
Figure 3.38: Obtained relevant values M^P and M^Q when the circuit is meshed by connecting nodes 54 and 94 for Scenario S1 (t_k = 640min)

beit in the opposite direction: the high M^P and M^Q values are found in lines along paths that link nodes with significant overvoltages and indicate areas of the network where power extractions could be beneficial control actions.

For other scenarios, the results in Appendix F reflect similar performance of



(a) Obtained M^P values



(b) Obtained M^Q values

Figure 3.39: Obtained relevant values M^P and M^Q when the circuit is meshed by connecting nodes 54 and 94 for Scenario S2 (t_k = 690min)

the metric when the system topology is changed or when there is additional power compensation from capacitor banks. The main conclusion drawn from the application of these metrics to each operational condition is that the metrics are capable of detecting the presence of controlled devices and quantitatively reflecting their

impact. This is particularly useful in cases where there is no prior knowledge of the system topology or when an assessment of the current system is desired, regardless of the installed devices or the system's topology. The metrics provide a value that indicates whether additional compensation is required when elements such as capacitor banks are installed, even if they are installed in different phases. Furthermore, in the case of a meshed distribution system, the metrics capture the characteristics of the topology and assess whether the newly formed paths, resulting from connections, are significant for modelling and control. This is possible because the values of voltages and power can reflect these changes in the system. It is conceivable that the proposed metrics can be calculated from time-series data collected following power injections performed at different locations in the network, which would facilitate the identification of nodes that have the greatest potential to impact the system voltage. As a result, these nodes can be considered as relevant input points from a control perspective.

3.5 Output analysis: Inferring the network state from observed voltages

Section 3.4 explored the characterisation of inputs—effective nodes for power injections or extractions—from information available in power through path connecting nodes and their connectivity. In this section it is explored the characterisation of possible *outputs*—effective nodes for voltage measurements—by studying the voltage variations and propagations caused by a perturbation in the system. Considering that the aim is to maintain a satisfactory voltage profile, the question that arises is which nodal voltages are the critical ones in the network, such that if these are measured and regulated, it can be inferred that network voltages *as a whole* are satisfactory.

The underlying idea behind these developments is the concept of *electrical distance* between nodes. The correlation between two nodal voltages is related to their electrical distance. If a voltage variation is observed at node i, the variation at node j can be inferred, albeit with some level of uncertainty, based on their electrical distance. When characterising possible sites for voltage measurement, it may be sufficient to measure the voltage at just one node among a collection of electrically close nodes, from which the voltages of the rest can be inferred. The traditional measure of electrical distance, calculated from sensitivities, quantifies the level of uncertainty and allows for the tracing of voltage variations across electrically close nodes in the network [107]. Similar ideas for modelling unbalanced distribution systems have been extensively presented in previous works [226–228]

for balanced systems and [216, 229–233] for unbalanced systems. These works analyse voltage sensitivity to detect the dominant factors contributing to voltage fluctuation. However, these approaches require prior knowledge of the system topology. Therefore, the challenge lies in estimating the electrical distances between different nodes in the network using measured data.

3.5.1 Definitions of electric distance

The electric distance can be used in order to measure the voltage variation propagated into the distribution system. Therefore, setting a definition of electric distance is required. A first definition corresponds to the impedance that is seen between nodes. The "closer" a node is from other, the smaller the impedance seen between nodes. This assumption is done in some papers, where an electrical distance is required to be calculated [215, 234, 235]. The sensitivity matrices presented before are used to get the electric distance by relating them to the distance between nodes and summarising the effect in the electric parameters associated with the electric line between two nodes. However, this assumption considers that all cables around the distribution system are the same (i.e., same material and same cross-section), which is not necessarily true in real systems.

A more precise discussion about electrical distance is done in [107, 236], in which different ways of defining any distance $D(i, j)$ are analysed under the verification of the following properties:

- Symmetry: $D(i, j) = D(j, i)$
- Posivity: $D(i, j) \geq 0$
- Nullity: $D(i, j) = 0 \Leftrightarrow i = j$
- Triangular inequality: $D(i, j) + D(j, k) \geq D(i, k)$

There are different ways to describe the relationship between nodes, e.g. using the sensitivity matrices R_{ij} and X_{ij} from equation (3.8), respectively, or obtaining the Z_{bus} from the system. The magnitude of coupling in terms of voltage between two nodes of a distribution system can be quantified by the maximum attenuation of voltage variation obtained from the previous matrices mentioned before. In general terms, a matrix of attenuation between all nodes α_{ij} is then available. Each term α_{ij} of the matrix gives a measure of the attenuation at node i with a voltage variation ΔV_i of a disturbance created at node j with a voltage variation ΔV_j. Equations (3.21) and (3.22) explains this relationship:

$$\Delta V_i = \alpha_{ij} \Delta V_j \tag{3.21}$$

where α_{ij} can be defined as follows:

$$\alpha_{ij} = \left| \frac{Z_{ij}}{Z_{jj}} \right| = \frac{\frac{\delta V_i}{\delta P_j} + \frac{\delta V_i}{\delta Q_j}}{\frac{\delta V_j}{\delta P_j} + \frac{\delta V_j}{\delta Q_j}} \tag{3.22}$$

It is important to remark that this matrix attenuation is not symmetric, since $\alpha_{ij} \neq \alpha_{ji}$. Additionally, a product of attenuation is required before changing over from a couple of nodes to others. Figure 3.40 illustrates this idea.

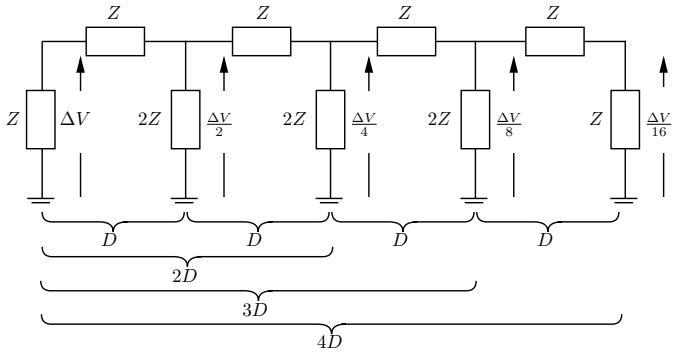


Figure 3.40: Example of voltage attenuation, which is affected by the electric distance seen in between. A longer distance will produce bigger voltage variation

An interpretation of α_{ij} after the electric distance definition can be the attenuation between nodes when the current flows from node j to node i. If the complex impedance $\tilde{Z}_{ij}$, $\tilde{Z}_{ji}$, $\tilde{Z}_{ii}$ and $\tilde{Z}_{jj}$ are considered and $\tilde{Z}_{ij} = \tilde{Z}_{ji}$, the following relationship can be presented:

$$\frac{\tilde{Z}_{ij}}{\tilde{Z}_{jj}} = \alpha_{ij} e^{i(\theta_{ij} - \theta_{jj})} = \alpha_{ij} e^{i(\Delta\theta_{ij})} \tag{3.23}$$

$$\frac{\tilde{Z}_{ji}}{\tilde{Z}_{ii}} = \alpha_{ji} e^{i(\theta_{ji} - \theta_{ii})} = \alpha_{ij} e^{i(\Delta\theta_{ji})} \tag{3.24}$$

It is possible to take the logarithm of attenuation as a definition of the distance between two nodes. Additionally, the formulation can consider both terms α_{ij} and α_{ji} to make the distance definition symmetric. Therefore, the electric distance can be written, as shown in Equation (3.25).

3.5.2 Results after evaluating definitions of electric distance

In[36], a detailed discussion about the different definitions of electric distance was presented, including the relationship with the voltage propagation around the distribution system. Some simulations are presented to validate these definitions

and their effect on the voltage outcome by comparison. For electric distance, two definitions are going to be compared: the first definition, which is based in electric impedance between nodes; the second definition corresponds to the following expression:

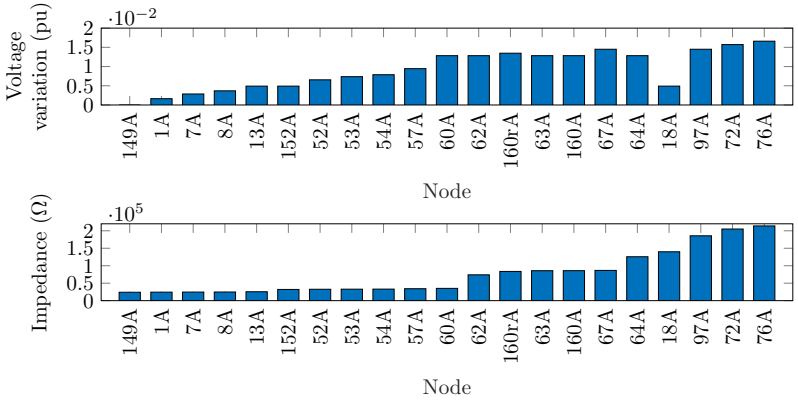
$$\begin{aligned} D(i,j) = D(j,i) &= -\ln\left(\frac{\tilde{Z}_{ij}^{\,2}}{\tilde{Z}_{ii}\tilde{Z}_{jj}}\right) \\ &= -\ln(\alpha_{ij}\alpha_{ji}) - i(\Delta\theta_{ij} + \Delta\theta_{ji}) = D_V(i,j) + D_{ph}(i,j) \end{aligned} \tag{3.25}$$

It is shown that electrical distance comprises a real component ($D_V(i,j)$) known as voltage electrical distance and an imaginary component ($D_{ph}(i,j)$) known as phase electrical distance. All the mathematical properties mentioned earlier can be proven for the obtained expression [107]. Since this analysis focuses solely on voltage variations, the concept of voltage electrical distance will be discussed. A mathematical expression for the voltage electrical distance between all nodes in the system can be defined using various elements of the network matrix.

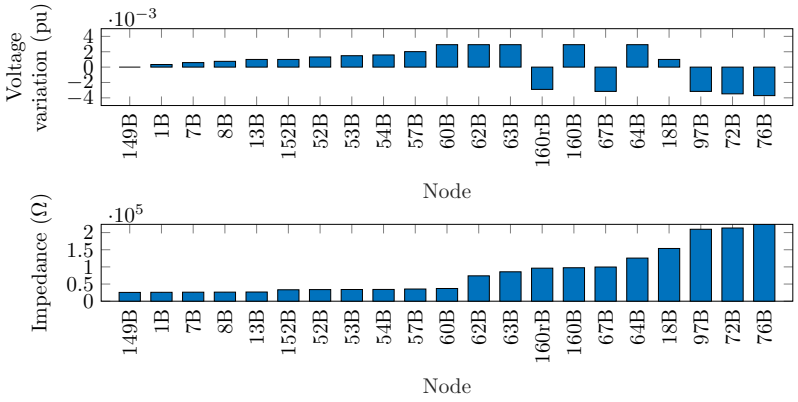
Figure 3.41 and Figure 3.42 present examples of nodes sorted by proximity using different definitions of electrical distance when a significant perturbation is applied to node 85, phase C, according to scenario S1. A comparison is made between nodes sorted by the impedance between nodes as a measure of electrical distance (as shown in Figure 3.41) and the voltage distance metric in equation (3.25) (as shown in Figure 3.42). In each case, the actual voltage variations observed at the same nodes are shown as a basis for comparison. The electrical distances are calculated with respect to the perturbation node, 85C; therefore, the distance indicated for 85C is zero, and the closest node (according to both distance definitions) is 84C. Tabulated values for this example and all simulated cases are presented in Appendix G.

In a similar way, Figure 3.43 and Figure 3.44 present examples of nodes sorted by proximity using different definitions of electrical distance when a significant perturbation is applied to node 85, phase C, according to scenario S2.

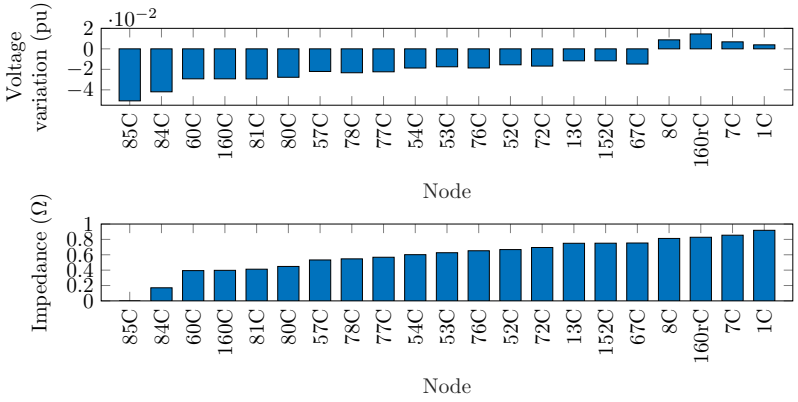
As shown in both examples, a positive voltage variation indicates an increase in voltage during the event, while a negative voltage variation reflects a decrease. The variation pattern differs for each phase, and it is determined by the voltage-power balance observed at each node. Since the perturbation is applied to phase C in both scenarios, the highest voltage variation is observed in the same phase. The variation follows the nature of the perturbation: in Scenario S1, the voltage on the same phase tends to decrease as it gets closer to the perturbed node due to the high power consumption, whereas in Scenario S2, the variation is positive as



(a) Phase A



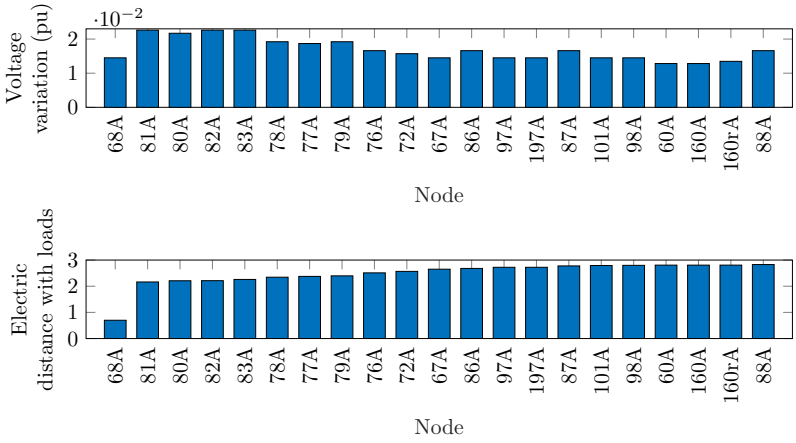
(b) Phase B



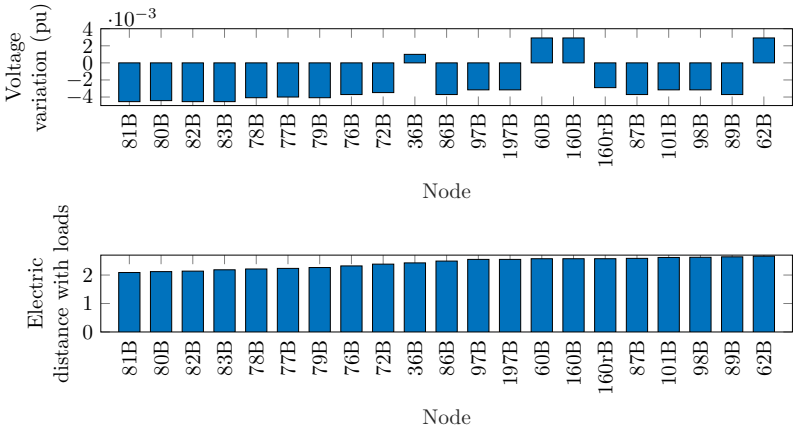
(c) Phase C

Figure 3.41: Electric distances and sorted voltages by measuring only impedance between nodes when there is a high perturbation at node 85, phase C (scenario S1, t_k = 640min)

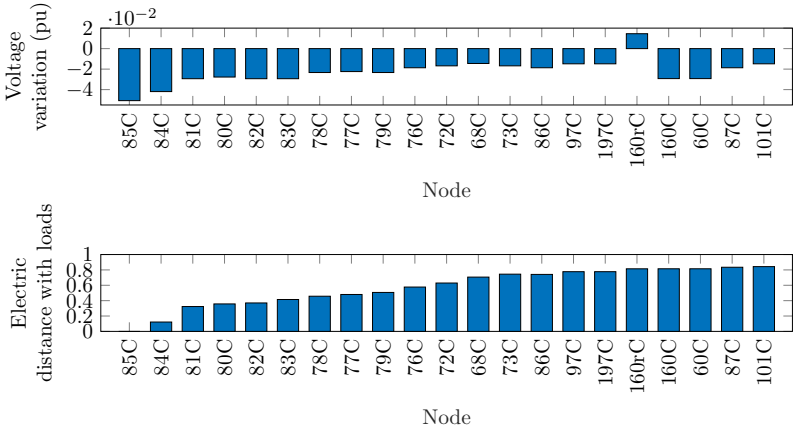
some power flows back to the feeder. Since the perturbation is only applied to one phase, there will be another phase that responds in the opposite manner as it gets closer to the perturbed node. Finally, the third phase acts to compensate for the



(a) Phase A

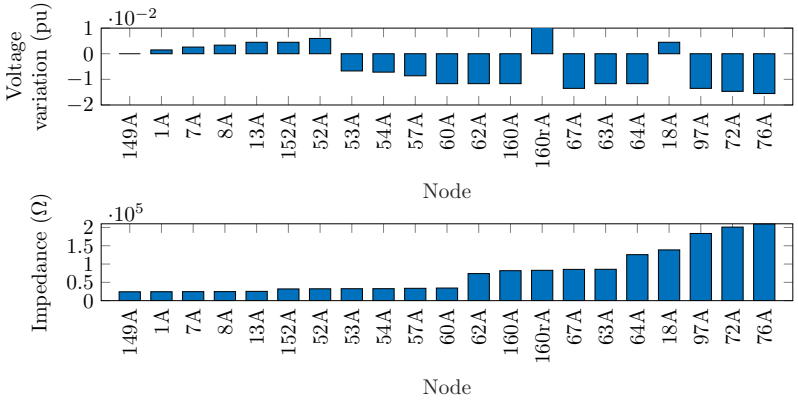


(b) Phase B

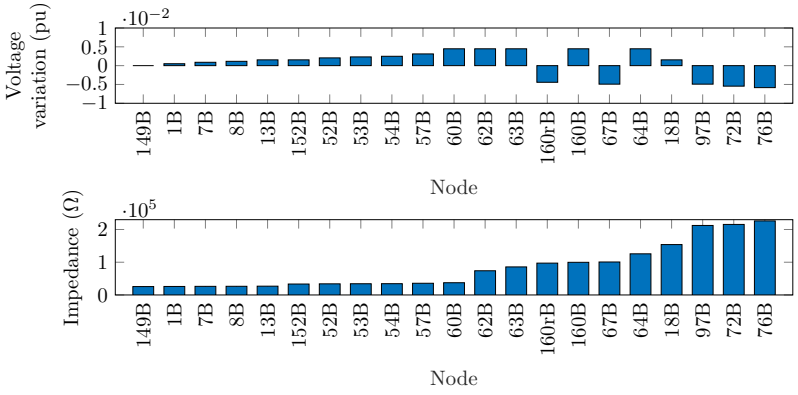


(c) Phase C

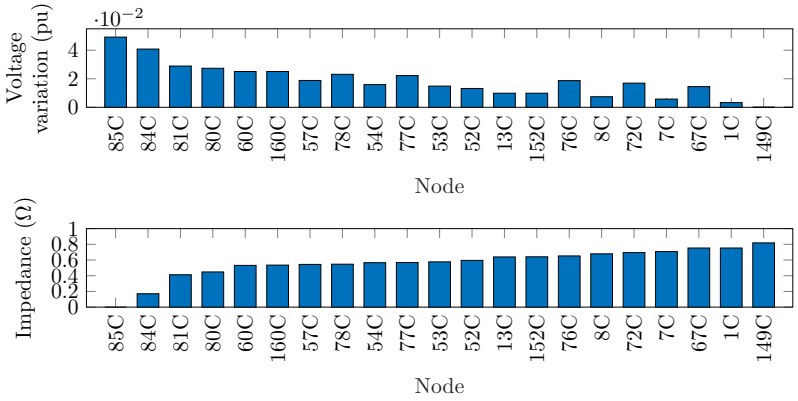
Figure 3.42: Voltage electrical distance and sorted voltages according to (3.25) when there is a high perturbation at node 85, phase C (scenario S1, t_k = 640min)



(a) Phase A



(b) Phase B

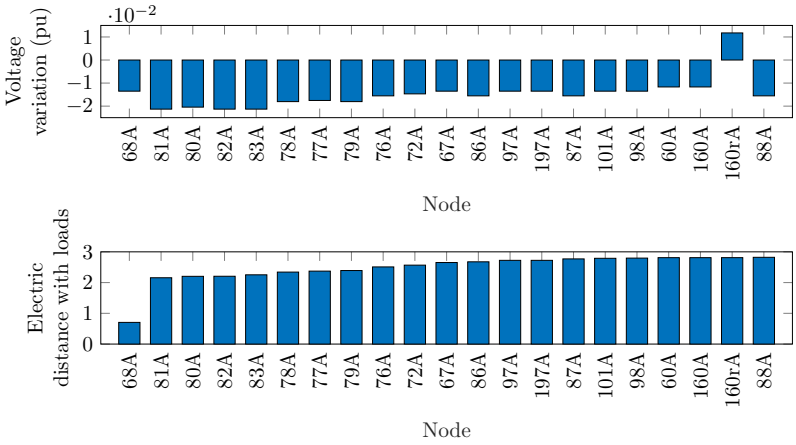


(c) Phase C

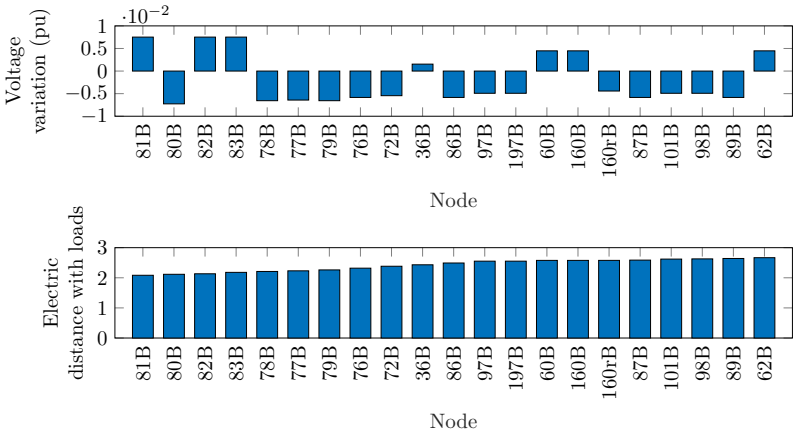
Figure 3.43: Electric distances and sorted voltages by measuring only impedance between nodes when there is a high perturbation at node 85, phase C (scenario S2, t_k = 690min)

effects observed in the other two phases, but the voltage variation (positive and negative) is not as significant as that seen in the other two phases.

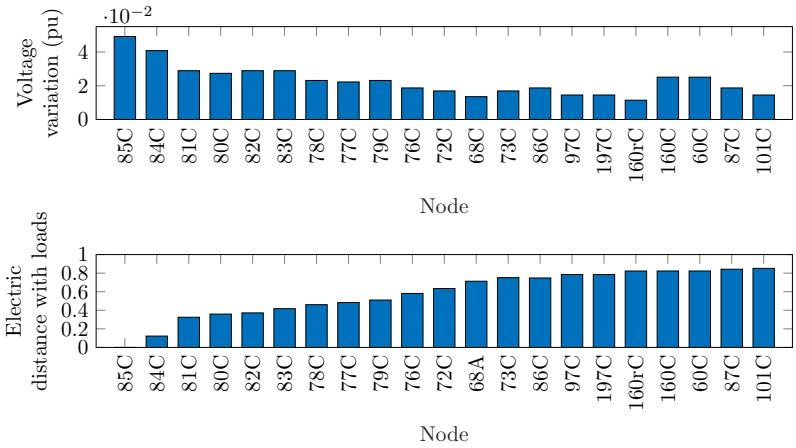
For the same event, voltage variations are sorted in different ways to gain a bet-



(a) Phase A



(b) Phase B



(c) Phase C

Figure 3.44: Voltage electrical distance and sorted voltages according to (3.25) when there is a high perturbation at node 85, phase C (scenario S2, t_k = 690min)

ter understanding of how the concept of distance can assist in building a model of voltage variation propagation. In the first example, Figure 3.41 displays voltages sorted according to the definition of impedance observed using the perturbation node as a reference (S85.C), while Figure 3.42 utilises the electrical distance presented in equation (3.25). The results obtained from both models indicate that both electrical distance measures provide a reasonable prediction of voltage variation: smaller electrical distances generally correspond to larger voltage variations after perturbation when the reference for calculation is over the same phase. Conversely, an inverse effect is observed over the other phases, as the same "point" would be the farthest if the reference is a fixed node within a specific phase. The differences in the ranked orders can be explained by the information considered in each electrical distance calculation: the impedance method focuses on the magnitudes of voltage variations but does not properly account for the network's topology, whereas the voltage electric distance metric considers both the voltage variation and the system's topology. For example, node 60C illustrates this distinction: it ranks third (in Scenario S1) when sorted by impedance-based electrical distance, but only 19th when sorted by the voltage electrical distance metric.

Figure 3.22 shows that the voltage variation at node 60C ranks as the 14th largest in the system following the perturbation. However, node 60C is located on the other side of a voltage regulator from the perturbation at 85C, which leads to a misleading effect on the impedance calculation. The voltage distance metric is more effective in filtering out this effect and accurately determining the nodes that are truly electrically close. Similar results were obtained for Scenarios S3, S4, and S5.

The impedance method primarily focuses on the magnitude of voltage variation, which does not necessarily reflect the reality of the system's topology. On the other hand, the voltage electric distance more accurately sorts the nodes by considering the voltage variation, system position, and topology. This information is crucial when clustering nodes is necessary for developing the system model for control purposes. Only nodes that are electrically close should be clustered together into a single 'node' where voltage is measured. In practice, voltage is measured at one of the nodes in the cluster, and the voltages at other nodes within the cluster can be inferred from the electrical distance.

The closest voltage variation can be modelled with topology knowledge and maximum variation and behaves in the same way as the observed system. Assuming that all nodes are measured, and the biggest voltage variation is detected, the system can be modelled according to equation (3.21). Figure 3.45 shows the results obtained from perturbation S1 as an example to illustrate this idea. Blue

bars correspond to the modelled voltage variations, while the rose (darken) bars represent the measured voltage variation. Perturbations S2-S5 can be modelled similarly. Therefore, it is possible to estimate the voltage variations in other nodes close enough in the sense of electrical distance, if voltage variations are observed in any node. It is possible to devise a concept of structural observability of proximity for which the electrical distance gives a quantified measurement of this concept.

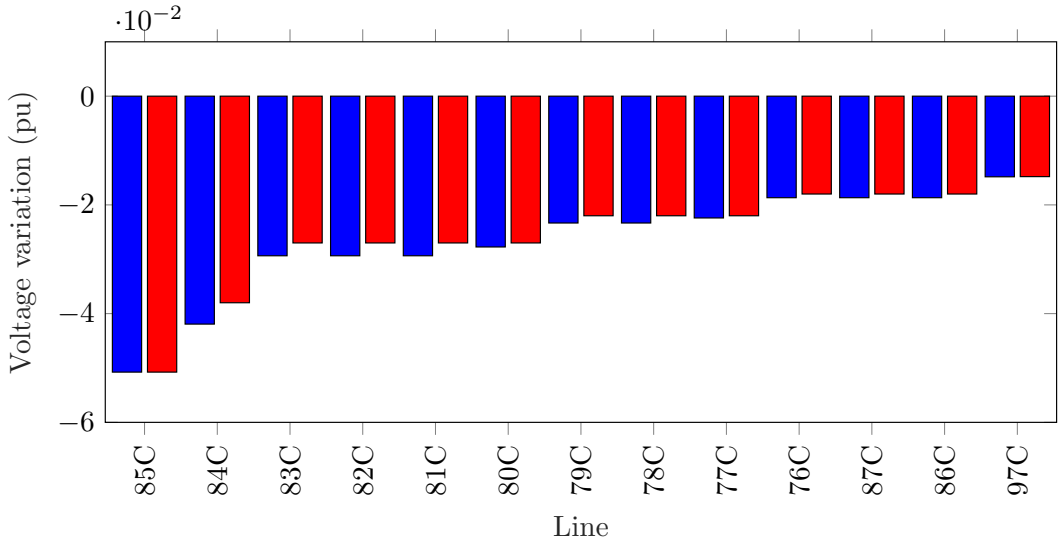


Figure 3.45: Voltage variation obtained from reference measurements (blue bars) and modelled using equation (3.21), for perturbation S1

It is shown from this figure that the closest voltage variation can be modelled with knowledge of topology and maximum variation and behaves in the same way than the system that is observed.

Similarly, when considering the nodes that control the system, the actuators cannot affect those distant nodes from the generating sets. They will have a marked influence only in its close vicinity (in this case, measured by the voltage attenuation). A concept of structural controllability of proximity can be devised from the electrical distance, which provides a quantified measurement of this concept (criteria to develop a clustering in the system, instead of modelling the complete distribution system).

The only case in which response was slightly different was scenario S6, in which two synchronised perturbations were done at nodes 66 and 85, both phase C. Figures 3.46 and 3.47 and showed the obtained sorted nodes with both electric distance concepts. and Figure 3.48 shows the obtained voltage attenuation model.

It can be seen that the sorting of nodes by the voltage distance metric in equation (3.25) identifies a set of nodes that are electrically close to 66C but that does

not extend to the branch that includes 85C; the impedance-based calculation, on the other hand, includes nodes in both sections of the network—those close to 66C and those close to 85C. It can be concluded that for two perturbations that occur simultaneously it would be difficult to identify electrically close nodes to each perturbation without further analysis.

Finally, the same analysis is repeated for the other scenarios, including cases where capacitor banks are connected or parts of the system are meshed. The results, presented in Appendix I, demonstrate that the addition of new devices such as capacitors does not drastically change the concept of using electrical distances, but it does impact the voltage variation. Moreover, the presence of a meshed circuit affects the distribution of power throughout the system, and consequently, the propagation of voltage in the distribution system. Nonetheless, the principle of using electrical distances still applies in a similar manner as previously introduced.

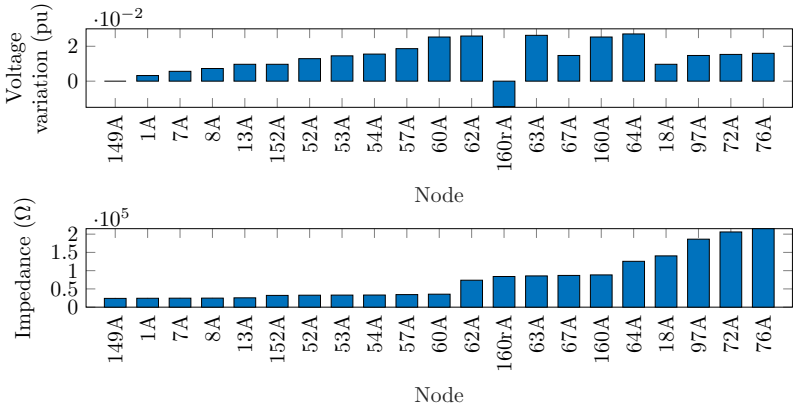
3.5.3 Use of covariance of voltage measurements

Section 3.5.2 introduced the electrical distance to develop a model that explains the voltage variation when perturbation in the distribution system. Unfortunately, quantifying the electrical distance requires real-time determination of impedances or voltage sensitivities at multiple discrete locations throughout the network [107]. As a practical alternative, therefore, it is sought a proxy for electrical distance that can be computed from available voltage measurements yet provide a similar insight. The main question is: how this electrical distance can be estimated when there is no knowledge from system topology? Is there any similar concept that can be used and gives a similar insight of electrical distance obtained from measurement?

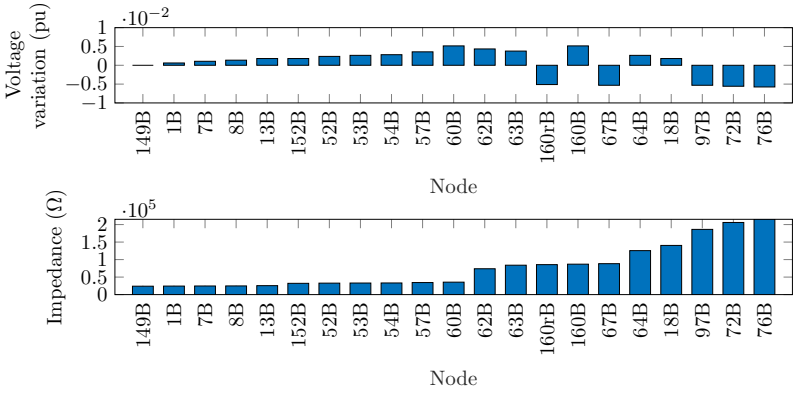
For this it is proposed to use the covariance between nodal voltages, which measures the joint variability of the random variables represented by the time-series voltage measurement [237]. The covariance sign shows the linear relationship between the variables, representing the distinction between voltage phases. For a pair of time-series voltage measurements (V_i, V_j) (representing two jointly distributed real random variables with finite second moments), the covariance is given by the expected value of the product of their deviations from their expected values:

$$\operatorname{cov}\left(V_i, V_j\right) = \mathrm{E}\left[\left(V_i - \mathrm{E}\left[V_i\right]\right)\left(V_j - \mathrm{E}\left[V_j\right]\right)\right] \tag{3.26}$$

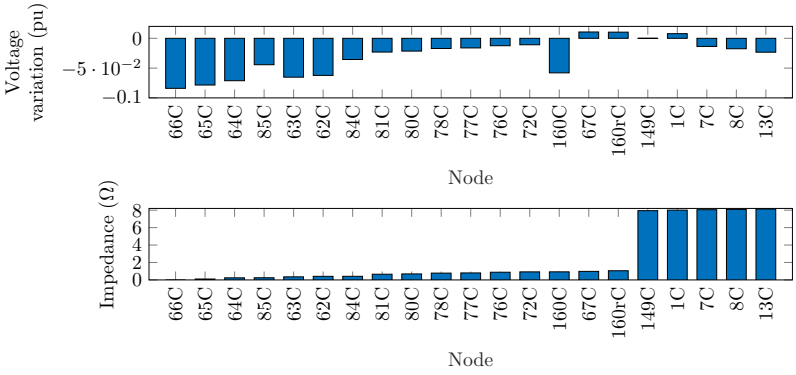
where $\mathrm{E}[V_i]$ and $\mathrm{E}[V_j]$ are the expected values of V_i and V_j, respectively. A covariance matrix Σ (known as dispersion matrix or variance–covariance matrix) is



(a) Phase A



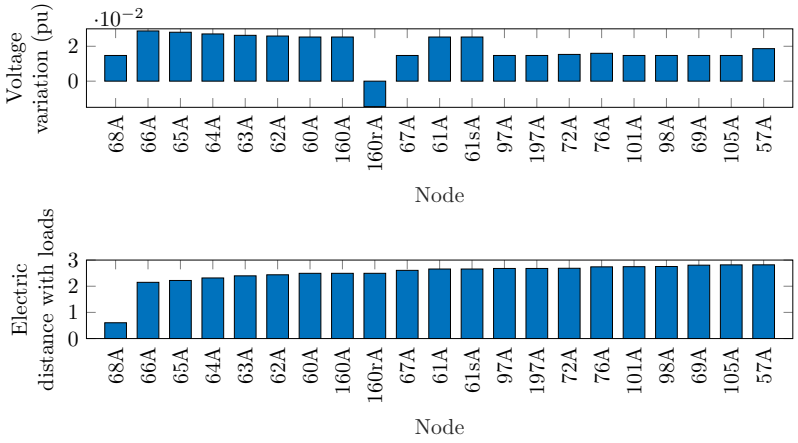
(b) Phase B



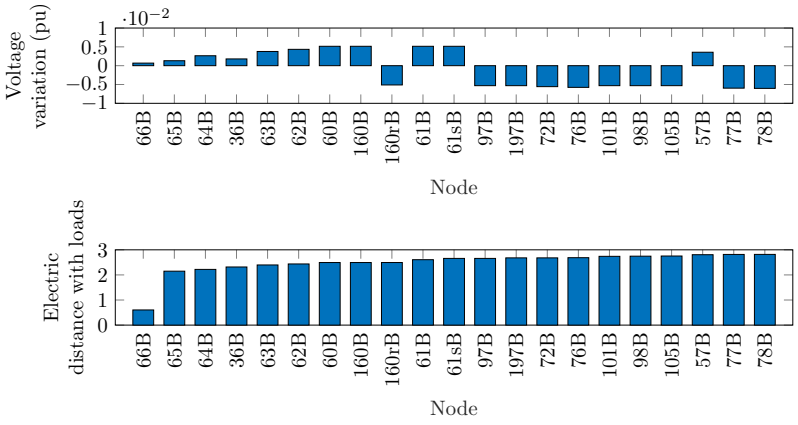
(c) Phase C

Figure 3.46: Electric distances and sorted voltages by measuring only impedance between nodes when there is a high perturbation at node 85, phase C (scenario S6, t_k = 690min)

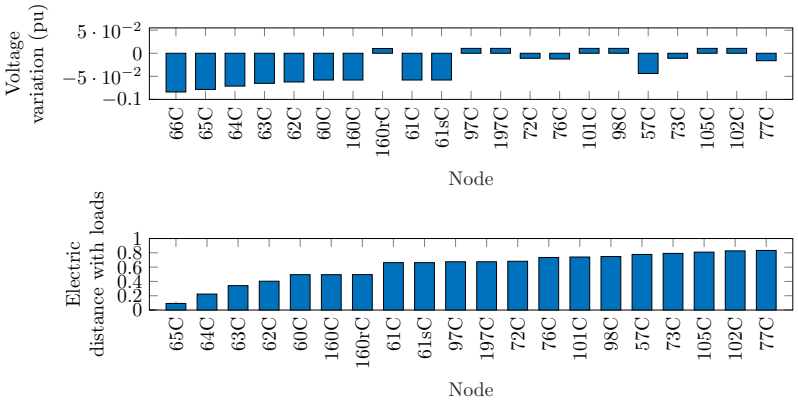
obtained from computing the covariance between each nodal measurement pair. For the measurement points $\{1, 2, \ldots, i, \ldots, j, \ldots, n\}$, the (i, j) entry of the covariance matrix Σ is the covariance between voltages V_i and V_j, where the (i, j) element



(a) Phase A



(b) Phase B



(c) Phase C

Figure 3.47: Voltage electrical distance and sorted voltages according to (3.25) when there is a high perturbation at node 85, phase C (scenario S6, t_k = 690min)

of the matrix is

$$\Sigma_{i,j} = \text{cov}\left(V_i, V_j\right). \tag{3.27}$$

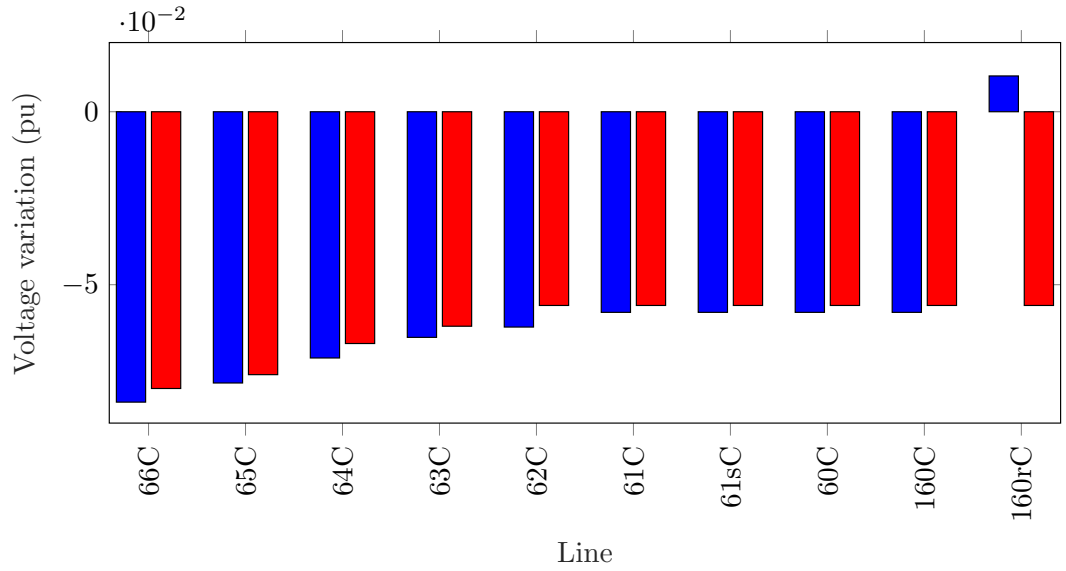


Figure 3.48: Voltage variation obtained from reference measurements (blue bars) and modelled using equation (3.21), for scenario S6

The matrix Σ is square, symmetric, positive semi-definite, and its diagonal contains variances (i.e., the covariance of each element with itself)–thus meets the axioms of a valid electrical distance metric [107]—and its diagonal contains variances (i.e., the covariance of each element with itself). Column j of Σ gives the covariance between the voltage measurement at node j and each of the other measured nodes.

The use of voltage correlations helps in describing the connectivity of the nodes across the different phases. It is proposed to use the voltage covariance to check how the propagation of the voltage variation moves around the system, which measures the joint variability of two or more random variables (represented by time-series voltage) [237].

Moreover, $\Sigma_{i,j}$ can be directly expressed as [108]:

$$\begin{aligned}\operatorname{cov}(V_i, V_j) = \mathrm{E}[(R_i(P_i - \mathrm{E}[P_i]) + (X_i(Q_i - \mathrm{E}[Q_i])) \\ (R_j(P_j - \mathrm{E}[P_j]) + (X_j(Q_j - \mathrm{E}[Q_j]))].\end{aligned} \quad (3.28)$$

Therefore, Σ contains information regarding the topology and impedance of the grid (encoded in R_i, R_j, X_i and X_j), further suggesting its suitability as a proxy measure of electrical distance.

An interesting advantage of this approach is that it provides a better understanding of the interaction between two nodes, as reflected by the covariance value along a path. For example, considering the configuration shown in Figure 3.29,

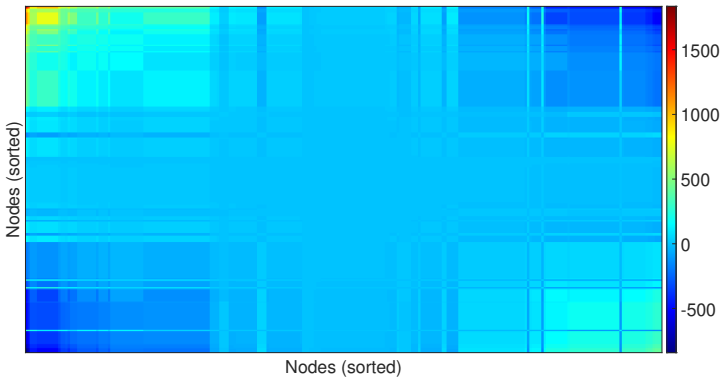


Figure 3.49: Covariance surface obtained from voltage measurements when there is a high perturbation at node 85, phase C (scenario S1, t_k = 640min)

the total power between two non-measurable points may not change significantly, but the equivalent resistance and reactance between the nodes would affect the obtained value, compared to the case where the device is not installed. This helps gain insight into the elements present within a path, even if the electrical information of the device is not available. Similarly, in the case presented in Figure 3.30, the covariance will also change based on the power variation along the path between two nodes. In summary, since covariance considers the equivalent impedance values and power balance between two nodes in a path, any component that can alter this balance will be reflected in the covariance value, thus providing relevant information about the current state of the network.

The calculation of covariance values presented in this section are tabuled in Appendix H. Figure 3.49 shows some of the biggest the values generated by matrix $\mathbf{\Sigma}$ when there is only a big perturbation at node 85, phase C, according to scenario S1 (load consumption). The higher range of variation in the covariance values is shown in the nodes close to the perturbation node. However, the absolute magnitude of the covariance obtained in this way lacks physical meaning and insight, and therefore a normalisation procedure is proposed by dividing for the biggest covariance value detected on each node to allow comparison of the columns or rows of Σ and, ultimately, identification of electrically close nodes. Figure 3.50 illustrates the results after the (raw) values are normalised. Table 3.3 shows the corresponding numerical values of the covariances and normalised covariances with respect to node 85C; that is, the column of Σ and its normalised counterpart corresponding to node 85C. Nodes close to the perturbation point show similar results.

It is important to highlight, in Table 3.3, that the largest covariance values

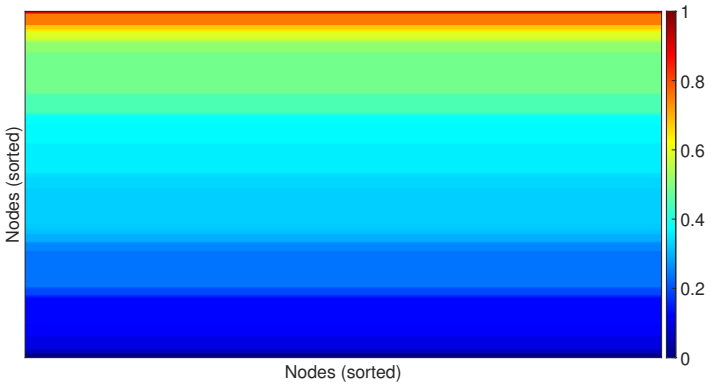


Figure 3.50: Normalised covariance surface obtained from voltage measurements when there is a high perturbation at node 85, phase C (scenario S1, t_k = 640min)

Table 3.3: Voltage covariance values obtained from a large perturbation at node 85, phase C (scenario S1).

Measured node (Node 85, Phase C)			Measured voltage	
$\Sigma_{i,j}$	Norm. $\Sigma_{i,j}$	Sorted node	Ranked nodes	Variation
1835.11	1.00	85, phase C	85, phase C	-5.08%
1490.11	0.87	84, phase C	84, phase C	-4.19%
1148.20	0.74	64, phase C	83, phase C	-2.94%
1148.20	0.74	65, phase C	81, phase C	-2.94%
1148.19	0.74	66, phase C	82, phase C	-2.94%
1148.19	0.74	160, phase C	64, phase C	-2.93%
1148.19	0.74	63, phase C	65, phase C	-2.93%
1148.19	0.74	61, phase C	66, phase C	-2.93%
1148.19	0.74	61, phase C	160, phase C	-2.93%
1148.19	0.74	62, phase C	63, phase C	-2.93%
1148.19	0.74	60, phase C	61, phase C	-2.93%
999.17	0.69	83, phase C	61, phase C	-2.93%
999.17	0.69	81, phase C	62, phase C	-2.93%

obtained are for nodes close to the perturbation point. Below the highest operational limit (in this case, 3% of rated voltage), the nodes are sorted according to this variation. Some of the presented nodes are close to the perturbation point, which is still contemplating the voltage electric distance presented before. From Figure 3.50, it is shown that the values of covariance obtained for each node are around the same range, which can give a consistent sense of position when they are compared.

Since these results only reflect the system's perception from a specific node, it is proposed to average the obtained normalised covariance over each node (*i.e.* each row of the normalised Σ is averaged across all columns). Table 3.4 shows the results of this for Scenario S1. The average normalised covariances higher than

Table 3.4: Average normalised covariance values obtained from a large perturbation at node 85, phase C (scenario S1).

Voltage variation	Rank nodes	Average norm. $\Sigma_{i,j}$	Rank nodes
−5.08%	85, phase C	0.99	85, phase C
−4.19%	84, phase C	0.86	84, phase C
−2.94%	83, phase C	0.75	64, phase C
−2.94%	81, phase C	0.75	65, phase C
−2.94%	82, phase C	0.75	66, phase C
−2.93%	64, phase C	0.75	160, phase C
−2.93%	65, phase C	0.75	63, phase C
−2.93%	66, phase C	0.75	61, phase C
−2.93%	160, phase C	0.75	61s, phase C
−2.93%	61, phase C	0.75	62, phase C
−2.93%	61s, phase C	0.75	60, phase C
−2.93%	62, phase C	0.67	83, phase C
−2.93%	63, phase C	0.67	81, phase C
−2.93%	60, phase C	0.67	82, phase C

0.85 correspond to nodes affected under the same voltage perturbation and can be considered the closest (and would be candidates for clustering as one node for the purpose of identifying a system model and determining which nodal voltages should be measured). The nodes that still are close enough to be impacted correspond to those with normalised covariance higher than 0.65. There are some nodes—such as node 64, phase C—which are highly ranked even though they are spatially far from the perturbation point. This is an effect of the voltage regulator in between, which via the action of the tap changer is modifying the equivalent electrical distance. The voltage variation is an effect of the power balance in the node plus the electric distance from the perturbation analysed.

The use of this average normalised covariance approach will only have relevant meaning in these critical scenarios (when the voltage surpass the allowed limits). For a situation in which the perturbation (generation or consumption) is not high, this will not have any special meaning, because the covariance surface will not be remarked for any peaks and all the power between nodes will be totally balanced.

Since, in Table 3.4, the averaged covariances follow a similar rank order to that of voltage variation per phase, it is suggested that the voltage covariance provides information closely related to the electrical distance between nodes, and appears to be an acceptable proxy for the latter. Scenarios S2, S3, S4, and S5 showed a similar set of results; for example, Figure 3.51 illustrates the covariance matrix, which exhibit a similar distribution and pattern of values as for scenario S1. Therefore, the average normalised covariance provided an acceptable proxy for electrical distance in the event of a single perturbation of either power injection or

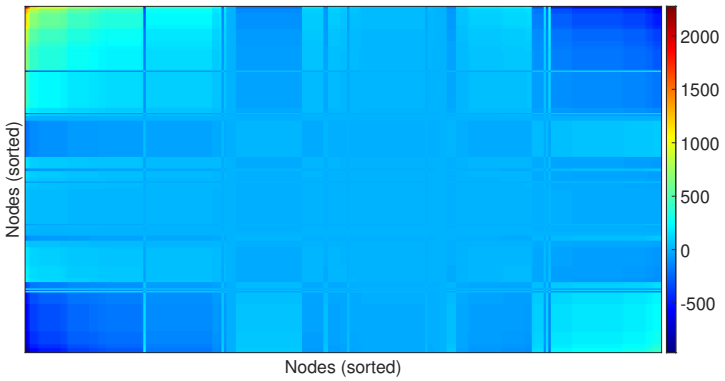


Figure 3.51: Covariance surface obtained from voltage measurements when there is a high perturbation at node 85, phase C (scenario S2, t_k = 690min)

extraction. The main advantage of this corresponds to a structure of the system's topology without previous knowledge, which is deducted from available voltage measurements. This procedure is still valid for different events e_{tk} (provided they are significant events), which will show similar results.

On the other hand, it is observed that for situations in which the perturbation (be it generation or consumption) is small or moderate (in the sense that voltages remained within bounds; scenarios that are milder than S1–S6), the covariance surface did not exhibit significant variations, essentially indicating balance within the system. The use of the average normalised covariance approach only has relevant meaning in *critical* scenarios (*i.e.*, when the voltage profile surpasses the operational limits).

To validate the covariance results, a Principal Component Analysis (PCA) was performed on the voltage measurements obtained under different scenarios. The largest eigenvalues for scenario S1 are presented in Table 3.5, shows that nearly 99% of the variance is explained by the first two principal components; details of the the corresponding eigenvectors are presented in Table 3.6.

Table 3.5: PCA for voltage measurements when a high perturbation occurs at node 85, phase C (scenario S1)

Number	Eigenvalue	Score
1	31598.98	89.29%
2	3022.70	8.54%
3	768.01	2.17%

The first eigenvector is sorted by using the weight of each component. The list and rank order of nodes is similar to that shown in Table 3.4, which were

Table 3.6: Sorted eigenvectors associated with the first two eigenvalues of the PCA for scenario S1

Eigenvector 1			Eigenvector 2		
Weight	Rank buses	Voltage variation	Weight	Rank buses	Voltage variation
0.24	85, phase C	-5.08%	0.15	160r, phase C	1.45%
0.20	84, phase C	-4.19%	0.14	105, phase C	-1.48%
0.15	64, phase C	-2.93%	0.14	108, phase C	-1.48%
0.15	65, phase C	-2.93%	0.14	67, phase C	-1.48%
0.15	66, phase C	-2.93%	0.14	197, phase C	-1.48%
0.15	160, phase C	-2.93%	0.14	97, phase C	-1.48%
0.15	63, phase C	-2.93%	0.14	100, phase C	-1.48%
0.15	61, phase C	-2.93%	0.14	104, phase C	-1.48%
0.15	61s, phase C	-2.93%	0.14	102, phase C	-1.48%
0.15	62, phase C	-2.93%	0.14	450, phase C	-1.48%
0.15	60, phase C	-2.93%	0.14	101, phase C	-1.48%
0.13	83, phase C	-2.94%	0.14	103, phase C	-1.48%
0.13	81, phase C	-2.94%	0.14	98, phase C	-1.48%

sorted using the average normalised covariance values. The second eigenvector obtained could not be explained using any apparent physical representation, and it is not providing any special information that can be used in this analysis. The top-ranked node is where there is a voltage regulator installed.

3.5.4 Impact of increasing the number of perturbations in the distribution system

To investigate the limitations of the proposed metrics, two simultaneous perturbations with the same time-series profile were performed at node 66 and 85, phase C (scenario S6). The observed voltage variations (ranked in descending order of magnitude) is presented in Figure 3.52, the corresponding surface from the covariance matrix is presented in Figure 3.53, and the summarised results with the normalised covariance values are shown in Table 3.7. The PCA test was again performed to compare the covariance analysis results, and the results of first eigenvalues are presented in Tables 3.8 and 3.9. Similar results to the previous case were obtained.

As observed with the electrical distance calculations for scenario S6, the ranking of nodes by voltage variations and average normalised covariances predominately determined by just one of the perturbations—the one with the highest voltage variation. The other perturbation effect is still presented in the sorted nodes, but it can not directly inform anything about the electrical distance; however, it is well known measuring the node impedance is not a good indicator of

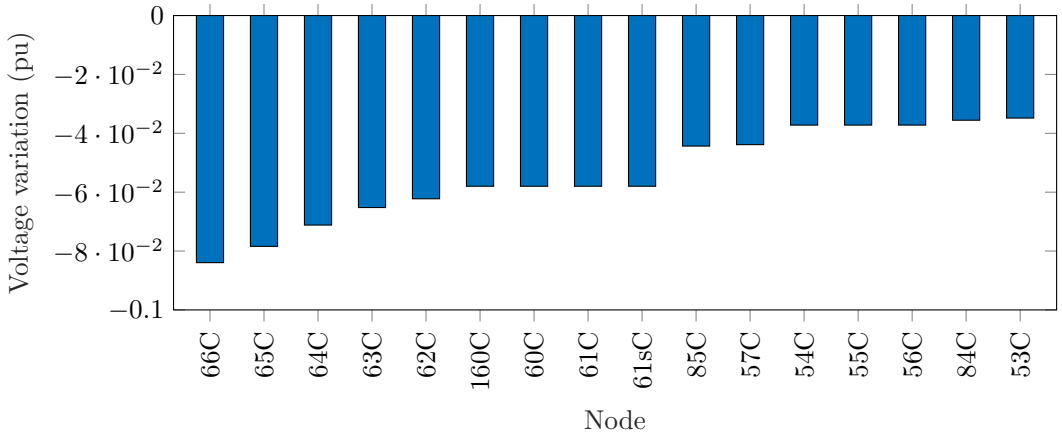


Figure 3.52: Voltage variation over each node when there are high perturbations at nodes 66 and 85, phase C (scenario S6, t_k = 640min)

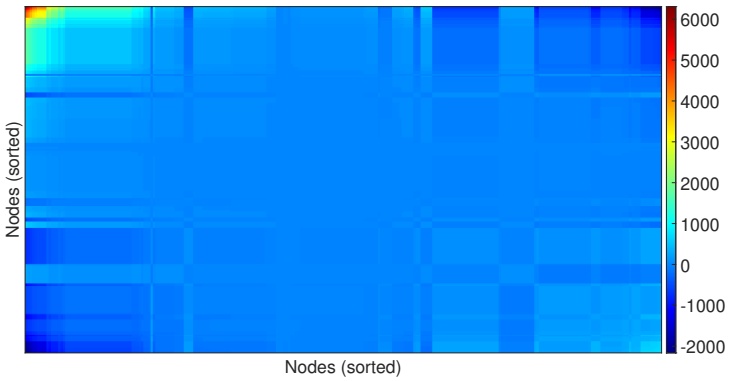


Figure 3.53: Covariance surface obtained from voltage measurements when there are high perturbations at nodes 66 and 85, phase C (scenario S6, t_k = 640min)

electrical distance when multiple perturbations occurs at the same time [107, 108]. The normalised covariance matrix roughly considers the nodes' electrical distance for the worst of the two perturbations. This means that a detailed analysis of the topology must be done, or previous information is required, to make concrete conclusions around electrical distances in the event of two perturbations.

3.5.5 Impact of reducing the number of measurements points

The last scenario for this methodology evaluated the performance to characterise the voltage when the number of measured nodes is reduced. Since the covariance matrix relies on the magnitude measured, the values and how values are sorted are the same for the fully observable case exposed before only if the maximum voltage variation is sensed in the measurement. The normalised values will use

Table 3.7: Average normalised covariance values obtained from high perturbations at nodes 66 and 85, phase C (scenario S6)

Voltage variation	Rank nodes	Average norm. $\Sigma_{i,j}$	Rank nodes
-8.39%	66, phase C	0.99	66, phase C
-7.84%	65, phase C	0.94	65, phase C
-7.12%	64, phase C	0.88	64, phase C
-6.52%	63, phase C	0.83	63, phase C
-6.22%	62, phase C	0.80	62, phase C
-5.80%	160, phase C	0.77	160, phase C
-5.80%	60, phase C	0.77	61, phase C
-5.80%	61, phase C	0.77	61s, phase C
-5.80%	61s, phase C	0.77	60, phase C
-4.43%	85, phase C	0.64	57, phase C
-4.39%	57, phase C	0.62	85, phase C
-3.72%	54, phase C	0.58	55, phase C
-3.72%	55, phase C	0.58	54, phase C
-3.72%	56, phase C	0.58	56, phase C
-3.56%	84, phase C	0.56	53, phase C
-3.48%	53, phase C	0.55	84, phase C
-3.10%	52, phase C	0.53	52, phase C
-2.33%	151, phase C	0.46	42, phase C
-2.33%	21, phase C	0.46	152, phase C

Table 3.8: PCA for voltage measurements when high perturbations occur at nodes 66 and 85, phase C (scenario S6)

Number	Eigenvalue	Score
1	76153.45	92.38%
2	5751.30	6.98%
3	528.01	0.64%

the same reference, a product from the vector with the highest voltage variation.

Sometimes this maximum variation cannot be measured, but some measuring units are installed in the surrounding nodes, which will impact the covariance matrix obtained. Nevertheless, the quasi-dynamics associated with the voltage variation could be still detected. To investigate this idea, the same analysis was applied for scenario S1 but reducing 50 measurement points (including the measurement at node 85, which is the node with the highest voltage variation). Tables 3.10 and 3.11 show the covariance values obtained and adjusted according to the available measurement. The PCA was also done to validate the obtained results, as shown in Tables 3.12 and 3.13.

The dynamic is still observed in the measurement is close enough to the perturbation point. The difference is that the nodes considered close to the perturbation will be referred according to the obtained normalised covariance values.

Table 3.9: Sorted eigenvectors from the first 2 eigenvalues of the PCA, case perturbations at nodes 66 and 85, phase C (scenario S6)

Eigenvector 1			Eigenvector 2		
Weight	Rank buses	Voltage variation	Weight	Rank buses	Voltage variation
0.299	66, phase C	-8.39%	0.137	160r, phase A	-1.46%
0.268	65, phase C	-7.84%	0.135	112, phase A	1.47%
0.243	64, phase C	-7.12%	0.135	108, phase A	1.47%
0.223	63, phase C	-6.52%	0.135	111, phase A	1.47%
0.213	62, phase C	-6.22%	0.135	197, phase A	1.47%
0.199	160, phase C	-5.80%	0.135	67, phase A	1.47%
0.198	61, phase C	-5.80%	0.135	71, phase A	1.47%
0.198	61s, phase C	-5.80%	0.135	97, phase A	1.47%
0.198	60, phase C	-5.80%	0.135	70, phase A	1.47%
0.150	57, phase C	-4.39%	0.135	100, phase A	1.47%
0.135	85, phase C	-4.43%	0.135	101, phase A	1.47%
0.127	55, phase C	-3.72%	0.135	113, phase A	1.47%
0.127	54, phase C	-3.72%	0.135	450, phase A	1.47%
0.127	56, phase C	-3.72%	0.135	68, phase A	1.47%
0.119	53, phase C	-3.48%	0.135	69, phase A	1.47%
0.106	52, phase C	-3.10%	0.135	99, phase A	1.47%
0.106	84, phase C	-3.56%	0.135	105, phase A	1.47%
0.080	42, phase C	-2.33%	0.135	109, phase A	1.47%
0.080	152, phase C	-2.33%	0.135	114, phase A	1.47%

Therefore, it will be slightly different from the actual connection scheme, but the model is still accurate enough. The only drawback from this methodology is the sensor's proximity to this perturbation point, which will require additional information if a complete analysis is required. Normally, sensors are placed in the system's critical points and can catch the most relevant voltage variation. These are complemented by the measurement of renewable energy units installed into the grid that increase the system's observability.

3.6 Validation of results

The final part of the simulation conducted in this chapter aimed to validate the consistency of the obtained metrics by comparing them with a reference "full model." In this case, identical simulations were performed using the same power profiles but with a higher data resolution of 1 minute. The purpose was to assess the differences compared to the measured values, which were assumed to be taken every 10 minutes. This analysis sought to demonstrate and confirm that, even in the partially observed case, all buses on the identified "path" are correctly identified, similar to the full case, and that the unobserved data closely resemble

Table 3.10: Voltage covariance values from a high perturbation at node 85, phase C (scenario S1), with reduced measurements

Measured node (Node 84, Phase C)			Measured voltage	
Covariance mag.	Normalised values	Sorted nodes	Ranked nodes	Variation
1213.76	1.00	84, phase C	84, phase C	-4.19%
919.47	0.85	65, phase C	83, phase C	-2.94%
919.47	0.85	66, phase C	81, phase C	-2.94%
919.47	0.85	63, phase C	65, phase C	-2.93%
919.47	0.85	61, phase C	66, phase C	-2.93%
919.47	0.85	62, phase C	61, phase C	-2.93%
820.52	0.79	83, phase C	62, phase C	-2.93%
820.52	0.79	81, phase C	63, phase C	-2.93%
769.83	0.77	80, phase C	80, phase C	-2.77%
693.78	0.73	57, phase C	78, phase C	-2.33%
631.26	0.69	78, phase C	77, phase C	-2.24%
602.83	0.68	77, phase C	57, phase C	-2.21%
588.08	0.67	54, phase C	54, phase C	-1.87%

Table 3.11: Av. normalised covariance values obtained from a high perturbation (scenario S1), with reduced measurement

Voltage variation	Rank nodes	Average norm. $\Sigma_{i,j}$	Rank nodes
-4.19%	84, phase C	0.99	84, phase C
-2.94%	83, phase C	0.87	65, phase C
-2.94%	81, phase C	0.87	66, phase C
-2.93%	65, phase C	0.87	63, phase C
-2.93%	66, phase C	0.87	61, phase C
-2.93%	61, phase C	0.87	62, phase C
-2.93%	62, phase C	0.78	83, phase C
-2.93%	63, phase C	0.78	81, phase C
-2.77%	80, phase C	0.75	80, phase C
-2.33%	78, phase C	0.74	57, phase C
-2.24%	77, phase C	0.68	54, phase C
-2.21%	57, phase C	0.68	55, phase C
-1.87%	54, phase C	0.68	56, phase C
-1.87%	55, phase C	0.67	78, phase C

Table 3.12: PCA for scenario S1 with reduced measurements

Number	Eigenvalue	Score
1	22716.84	88.47%
2	2364.542	9.21%
3	596.6255	2.32%

the observed data in the full case. To ensure comparability, the evaluation window for the data remained at 60 minutes, consistent with the previous analysis.

Table 3.13: Sorted eigenvectors from the first 2 eigenvalues of the PCA(scenario S1), with reduced measurements

Eigenvector 1			Eigenvector 2		
Weight	Rank buses	Voltage variation	Weight	Rank buses	Voltage variation
0.22983	84, phase C	-4.19%	0.17	66, phase C	-2.93%
0.17463	65, phase C	-2.93%	0.17	65, phase C	-2.93%
0.17463	66, phase C	-2.93%	0.17	61, phase C	-2.93%
0.17463	63, phase C	-2.93%	0.17	62, phase C	-2.93%
0.17463	61, phase C	-2.93%	0.17	63, phase C	-2.93%
0.17463	62, phase C	-2.93%	0.13	57, phase C	-2.21%
0.15528	83, phase C	-2.94%	0.11	54, phase C	-1.87%
0.15528	81, phase C	-2.94%	0.11	55, phase C	-1.87%
0.14566	80, phase C	-2.77%	0.11	56, phase C	-1.87%
0.13177	57, phase C	-2.21%	0.10	53, phase C	-1.75%
0.11951	78, phase C	-2.33%	0.09	52, phase C	-1.56%
0.114	77, phase C	-2.24%	0.07	152, phase C	-1.17%
0.1117	54, phase C	-1.87%	0.07	35, phase C	-1.17%

Consequently, the data used to construct the metrics consisted of 60 data points, as opposed to the 6 data points used for the 10-minute resolution.

The initial stage focused on the metrics derived from power injections. Tables from 3.14 to 3.19 summarize the differences in key values compared to the base case, where measurements were taken every 10 minutes. Validating the system with fewer installed measurement units for this power-related metric is not necessary, as it is challenging to compare the connectivity of nodes in large distribution systems. However, objective comparisons can be made for node voltage measurements by simply comparing the two cases with a resolution of 1 minute and measurements taken every 10 minutes.

The subsequent six tables present the results after scenarios S1, S2, and S6, where no capacitor banks are installed, and the system employs a radial topology. The numerical values of the obtained metrics differ from those of the reference case. This discrepancy arises because the metrics are highly influenced by the measured values and the events detected in those variables. The power flowing through the lines, which is captured in the measurements, significantly impacts the absolute values of the metrics M^P and M^Q. Nevertheless, the ordering of the lines remains consistent, and the proportions between the obtained values are quite similar in comparison to the system measured every 1 minute. Importantly, the identified path in both cases is exactly the same, which is the primary objective of these metrics.

The subsequent validation focuses on the propagation of voltage variation

Table 3.14: Validation for M^P values during the event analysed in scenario S1.

Nodes		M^P (1 minute)	Nodes		M^P (10 minute)	Difference
57.C	60.C	71.06	57.C	60.C	36.83	48.17%
52.C	152.C	41.57	52.C	152.C	21.35	48.63%
57.C	54.C	36.95	57.C	54.C	19.02	48.51%
7.C	1.C	32.63	7.C	1.C	16.69	48.86%
8.C	13.C	32.48	8.C	13.C	16.64	48.76%
64.C	65.C	25.16	65.C	64.C	14.4	42.58%
1.C	149.C	23.50	1.C	149.C	12.4	47.21%
8.C	7.C	22.80	63.C	64.C	11.9	47.75%
53.C	52.C	22.58	66.C	65.C	11.8	47.93%
64.C	63.C	20.76	8.C	7.C	11.7	43.87%
66.C	65.C	20.57	53.C	52.C	11.6	43.76%
84.C	81.C	17.57	62.C	60.C	8.62	50.96%
84.C	85.C	17.27	85.C	84.C	7.98	53.79%
62.C	60.C	15.04	84.C	81.C	7.98	46.93%
54.C	53.C	14.87	54.C	53.C	7.61	48.86%
63.C	62.C	11.31	62.C	63.C	6.45	42.98%
80.C	78.C	4.58	78.C	80.C	2.12	53.82%
77.C	76.C	3.22	77.C	76.C	1.51	53.10%
81.C	80.C	2.47	81.C	80.C	1.16	53.25%
72.C	67.C	1.86	67.C	72.C	0.93	50.08%
76.C	72.C	1.80	67.C	160r.C	0.91	49.36%
67.C	160r.C	1.68	72.C	76.C	0.87	48.48%
78.C	77.C	1.38	77.C	78.C	0.65	52.81%

across the system. Tables from 3.20 to 3.25 summarise the differences in key values compared to the base case, where measurements were taken every 10 minutes, and considering partial installation of measurement units. Similarly to the previous tables, the following six tables present the results after scenarios S1, S2, and S6, where no capacitor banks were installed, and the system adopted a radial topology.

The metrics obtained for the average normalised covariance exhibit slight differences in values between the reference case with measurements installed throughout the system and the partially observable cases. However, these differences consistently remain in terms of numerical values. A similar effect is observed for the weights obtained in the main two eigenvectors derived from PCA. It is noteworthy that the voltage variations captured every minute reflect larger changes compared to the model that only took measurements every 10 minutes, particularly in the case of S6. Nevertheless, for all cases, it is demonstrated that the nodes impacted follow the same pattern. The voltage levels across the nodes, which are captured in the measurements, significantly influence the absolute values of the average normalised covariance metric, as it is a relative value. Similarly to the previous findings, the sorting of nodes remains consistent, and the proportions

Table 3.15: Validation for M^Q values during the event analysed in scenario S1.

Nodes		M^Q (1 minute)	Nodes		M^Q (10 minute)	Difference
57.C	60.C	159.60	57.C	60.C	82.71	48.18%
52.C	152.C	95.92	52.C	152.C	49.27	48.63%
57.C	54.C	82.98	57.C	54.C	42.72	48.52%
7.C	1.C	75.28	7.C	1.C	38.50	48.86%
8.C	13.C	74.95	8.C	13.C	38.40	48.76%
1.C	149.C	54.23	1.C	149.C	28.6	47.22%
8.C	7.C	52.61	8.C	7.C	26.9	48.90%
53.C	52.C	52.11	53.C	52.C	26.7	48.78%
54.C	53.C	34.31	54.C	53.C	17.5	48.86%
84.C	81.C	17.81	85.C	84.C	8.09	54.56%
84.C	85.C	17.51	84.C	81.C	8.09	53.79%
64.C	65.C	12.47	65.C	64.C	7.17	42.54%
80.C	78.C	10.29	63.C	64.C	5.91	42.59%
64.C	63.C	10.29	66.C	65.C	5.84	43.28%
66.C	65.C	10.20	78.C	80.C	4.75	53.39%
62.C	60.C	7.45	62.C	60.C	4.27	42.69%
77.C	76.C	7.23	77.C	76.C	3.39	53.09%
63.C	62.C	5.61	62.C	63.C	3.2	42.95%
81.C	80.C	5.56	81.C	80.C	2.6	53.25%
72.C	67.C	4.19	67.C	72.C	0.93	77.78%
76.C	72.C	4.05	67.C	160r.C	0.91	77.46%
67.C	160r.C	3.78	72.C	76.C	0.87	77.06%
78.C	77.C	3.11	77.C	78.C	0.65	79.00%

Table 3.16: Validation for M^P values during the event analysed in scenario S2.

Nodes		M^P (1 minute)	Nodes		M^P (10 minute)	Difference
84.C	81.C	17.05	84.C	81.C	19.32	-13.31%
85.C	84.C	14.94	85.C	84.C	16.51	-10.51%
60.C	57.C	12.94	60.C	57.C	12.94	0.03%
52.C	152.C	7.60	152.C	52.C	7.57	0.34%
57.C	54.C	6.53	57.C	54.C	6.54	-0.10%
8.C	13.C	5.89	8.C	13.C	5.85	0.63%
7.C	1.C	5.82	7.C	1.C	5.81	0.16%
80.C	78.C	4.87	80.C	78.C	5.64	-15.86%
52.C	53.C	4.12	76.C	77.C	4.52	-9.82%
8.C	7.C	4.08	53.C	52.C	4.11	-0.75%
76.C	77.C	3.87	8.C	7.C	4.07	-5.24%
149.C	1.C	2.79	67.C	160r.C	3.32	-19.04%
160r.C	67.C	2.77	72.C	67.C	3.03	-9.23%
53.C	54.C	2.71	1.C	149.C	2.74	-1.32%
67.C	72.C	2.58	54.C	53.C	2.70	-4.37%
81.C	80.C	2.32	80.C	81.C	2.60	-12.18%
76.C	72.C	2.15	76.C	72.C	2.48	-15.05%
78.C	77.C	1.36	78.C	77.C	1.52	-11.88%

between the obtained values are quite similar in relation to the system measured every 1 minute. This alignment with the system measured at a higher resolution

Table 3.17: Validation for M^Q values during the event analysed in scenario S2.

Nodes		M^Q (1 minute)	Nodes		M^Q (10 minute)	Difference
60.C	57.C	29.08	60.C	57.C	29.07	0.03%
52.C	152.C	17.54	84.C	81.C	19.59	-11.67%
84.C	81.C	17.28	152.C	52.C	17.48	-1.13%
85.C	84.C	15.15	85.C	84.C	16.74	-10.51%
57.C	54.C	14.67	57.C	54.C	14.69	-0.10%
8.C	13.C	13.59	8.C	13.C	13.50	0.63%
7.C	1.C	13.44	7.C	1.C	13.42	0.16%
80.C	78.C	10.93	80.C	78.C	12.66	-15.86%
52.C	53.C	9.51	76.C	77.C	10.16	-6.86%
8.C	7.C	9.41	53.C	52.C	9.48	-0.74%
76.C	77.C	8.70	8.C	7.C	9.41	-8.17%
149.C	1.C	6.45	67.C	160r.C	7.47	-15.83%
53.C	54.C	6.25	72.C	67.C	6.8	-8.81%
160r.C	67.C	6.22	1.C	149.C	6.33	-1.72%
67.C	72.C	5.80	54.C	53.C	6.22	-7.27%
81.C	80.C	5.21	80.C	81.C	5.85	-12.18%
76.C	72.C	4.84	76.C	72.C	5.56	-15.05%
78.C	77.C	3.06	78.C	77.C	3.42	-11.89%

Table 3.18: Validation for M^P values during the event analysed in scenario S6.

Nodes		M^P (1 minute)	Nodes		M^P (10 minute)	Difference
57.C	60.C	20.51	60.C	57.C	10.27	49.91%
84.C	81.C	18.05	84.C	81.C	8.36	53.70%
85.C	84.C	17.26	85.C	84.C	7.99	53.70%
52.C	152.C	11.82	152.C	52.C	5.88	50.26%
54.C	57.C	10.51	57.C	54.C	5.23	50.20%
7.C	1.C	9.20	7.C	1.C	4.57	50.36%
8.C	13.C	9.20	8.C	13.C	4.56	50.37%
7.C	8.C	6.45	8.C	7.C	3.21	50.21%
53.C	52.C	6.42	52.C	53.C	3.19	50.33%
1.C	149.C	6.02	1.C	149.C	2.98	50.44%
78.C	80.C	4.89	80.C	78.C	2.36	51.69%
54.C	53.C	4.20	53.C	54.C	2.08	50.39%
77.C	76.C	3.62	77.C	76.C	1.82	49.77%
81.C	80.C	2.55	67.C	160r.C	1.25	50.94%
67.C	72.C	2.25	81.C	80.C	1.23	45.40%
67.C	160r.C	2.13	67.C	72.C	1.21	43.11%
76.C	72.C	2.05	76.C	72.C	1.06	48.46%
78.C	77.C	1.46	78.C	77.C	0.72	50.82%

represents the main objective of this metric.

Table 3.19: Validation for M^Q values during the event analysed in scenario S6.

Nodes		M^Q (1 minute)	Nodes		M^Q (10 minute)	Difference
57.C	60.C	46.05	60.C	57.C	23.07	49.91%
52.C	152.C	27.26	152.C	52.C	13.56	50.27%
54.C	57.C	23.60	57.C	54.C	11.75	50.20%
7.C	1.C	21.23	7.C	1.C	10.54	50.37%
8.C	13.C	21.21	8.C	13.C	10.53	50.37%
84.C	81.C	18.30	84.C	81.C	8.47	53.70%
85.C	84.C	17.50	85.C	84.C	8.10	53.70%
7.C	8.C	14.87	8.C	7.C	7.40	50.22%
53.C	52.C	14.80	52.C	53.C	7.35	50.34%
1.C	149.C	13.88	1.C	149.C	6.88	50.45%
78.C	80.C	10.98	80.C	78.C	5.30	51.69%
54.C	53.C	9.69	53.C	54.C	4.81	50.40%
77.C	76.C	8.14	77.C	76.C	4.09	49.77%
81.C	80.C	5.73	67.C	160r.C	2.81	50.95%
67.C	72.C	5.05	81.C	80.C	2.76	45.39%
67.C	160r.C	4.78	67.C	72.C	2.72	43.11%
76.C	72.C	4.60	76.C	72.C	2.37	48.46%
78.C	77.C	3.28	78.C	77.C	1.62	50.81%

Table 3.20: Validation for voltage variation and average normalised covariance in scenario S1.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)	
	Voltage variation	Average norm. $\Sigma_{i,j}$	Voltage variation	Diff. voltage variation	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$
S85.C	-6.69%	0.99	-5.08%	31.8%	0.99	0.0%	–	–
S84.C	-5.39%	0.85	-4.19%	28.6%	0.86	-1.2%	0.99	-14.0%
S66.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	0.87	-10.0%
S64.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	–	–
S65.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	0.87	-10.0%
S160.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	–	–
S63.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	0.87	-10.0%
S61.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	0.87	-10.0%
S61s.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	–	–
S62.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	0.87	-10.0%
S60.C	-4.43%	0.77	-2.93%	51.3%	0.75	2.6%	–	–
S83.C	-3.54%	0.65	-2.94%	20.6%	0.67	-3.1%	0.78	-13.0%
S82.C	-3.54%	0.65	-2.94%	20.6%	0.67	-3.1%	0.78	-13.0%
S81.C	-3.54%	0.65	-2.94%	20.6%	0.67	-3.1%	–	–
S57.C	-3.34%	0.66	-2.21%	51.4%	0.64	3.0%	0.75	-9.0%
S80.C	-3.30%	0.62	-2.77%	19.0%	0.65	-4.8%	0.74	-12.0%
S56.C	-2.83%	0.6	-1.87%	51.5%	0.59	1.7%	–	–
S54.C	-2.83%	0.6	-1.87%	51.5%	0.59	1.7%	0.67	-7.0%
S55.C	-2.83%	0.6	-1.87%	51.5%	0.59	1.7%	0.67	-7.0%
S53.C	-2.65%	0.58	-1.75%	51.5%	0.57	1.7%	–	–
S78.C	-2.65%	0.55	-2.33%	13.6%	0.58	-5.5%	0.68	-13.0%
S79.C	-2.65%	0.55	-2.33%	13.6%	0.58	-5.5%	0.68	-13.0%
S52.C	-2.36%	0.55	-1.56%	51.6%	0.55	0.0%	–	–
S77.C	-2.51%	0.54	-2.24%	12.2%	0.57	-5.6%	0.68	-14.0%

3.7 Practical implementation of the proposed metrics

From the previous results, it is evident that both metrics for power injections and voltage variations provide insights into the system condition without requiring

Table 3.21: Validation for eigenvector values in scenario S1.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)			
	Eigenv. 1	Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2
S85.C	0.24	-0.02	0.24	0.0%	-0.04	8.3%	–	–	–	–
S84.C	0.2	0.02	0.20	0.0%	0.00	10.0%	0.23	-14.92%	0.00	0.00%
S66.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	–	–	–	–
S64.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	0.17	-16.42%	0.17	221.43%
S65.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	0.17	-16.42%	0.17	221.43%
S160.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	0.17	-16.42%	0.17	221.43%
S63.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	–	–	–	–
S61.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	0.17	-16.42%	0.17	221.43%
S61s.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	–	–	–	–
S62.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	0.17	-16.42%	0.17	221.43%
S60.C	0.17	-0.12	0.15	11.8%	-0.14	13.3%	–	–	–	–
S83.C	0.13	0.07	0.13	0.0%	0.06	7.7%	0.16	-19.45%	-0.07	221.43%
S82.C	0.13	0.07	0.13	0.0%	0.06	7.7%	0.16	-19.45%	-0.07	221.43%
S81.C	0.13	0.07	0.13	0.0%	0.06	7.7%	–	–	–	–
S80.C	0.12	0.08	0.12	0.0%	0.07	8.3%	0.15	-21.38%	-0.08	218.18%
S57.C	0.13	-0.09	0.11	15.4%	-0.11	18.2%	0.13	-19.79%	0.13	218.18%
S78.C	0.09	0.1	0.10	-11.1%	0.09	10.0%	0.12	-19.51%	-0.11	218.18%
S79.C	0.09	0.1	0.10	-11.1%	0.09	10.0%	–	–	–	–
S77.C	0.09	0.1	0.10	-11.1%	0.10	0.0%	0.11	-14.00%	-0.12	222.22%
S56.C	0.11	-0.08	0.10	9.1%	-0.09	10.0%	0.11	-11.70%	0.11	222.22%
S54.C	0.11	-0.08	0.10	9.1%	-0.09	10.0%	0.11	-11.70%	0.11	222.22%
S55.C	0.11	-0.08	0.10	9.1%	-0.09	10.0%	0.11	-11.70%	0.11	222.22%
S53.C	0.1	-0.07	0.09	10.0%	-0.08	11.1%	0.10	-11.70%	0.1	225.00%

Table 3.22: Validation for voltage variation and average normalised covariance in scenario S2.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)		
	Voltage variation	Average norm. $\Sigma_{i,j}$	Voltage variation	Diff. voltage variation	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$	
S85.C	4.93%	0.99	4.92%	0.4%	0.99	0.0%	–	–	
S84.C	4.09%	0.85	4.08%	0.4%	0.87	-2.4%	0.99	-14.0%	
S82.C	2.90%	0.77	2.89%	0.4%	0.70	9.1%	0.87	-10.0%	
S81.C	2.90%	0.77	2.89%	0.4%	0.70	9.1%	–	–	
S83.C	2.90%	0.77	2.89%	0.4%	0.70	9.1%	0.87	-10.0%	
S80.C	2.75%	0.77	2.73%	0.4%	0.68	11.7%	–	–	
S66.C	2.51%	0.77	2.51%	0.3%	0.65	15.6%	0.87	-10.0%	
S64.C	2.51%	0.77	2.51%	0.3%	0.65	15.6%	0.87	-10.0%	
S63.C	2.51%	0.77	2.51%	0.3%	0.65	15.6%	–	–	
S65.C	2.51%	0.77	2.51%	0.3%	0.65	15.6%	0.87	-10.0%	
S160.C	2.51%	0.77	2.51%	0.3%	0.65	15.6%	–	–	
S62.C	2.51%	0.65	2.51%	0.3%	0.65	0.0%	0.78	-13.0%	
S60.C	2.51%	0.65	2.51%	0.3%	0.65	0.0%	0.78	-13.0%	
S61.C	2.51%	0.65	2.51%	0.3%	0.65	0.0%	–	–	
S61s.C	2.51%	0.66	2.51%	0.3%	0.65	1.5%	0.75	-9.0%	
S78.C	2.32%	0.62	2.31%	0.4%	0.62	0.0%	0.74	-12.0%	
S79.C	2.32%	0.6	2.31%	0.4%	0.62	-3.3%	–	–	
S77.C	2.23%	0.6	2.22%	0.4%	0.61	-1.7%	0.67	-7.0%	
S57.C	1.88%	0.6	1.88%	0.3%	0.56	6.7%	0.67	-7.0%	
S91.C	1.88%	0.58	1.87%	0.5%	0.56	3.4%	–	–	
S95.C	1.88%	0.55	1.87%	0.5%	0.56	-1.8%	0.68	-13.0%	
S92.C	1.88%	0.55	1.87%	0.5%	0.56	-1.8%	0.68	-13.0%	
S93.C	1.88%	0.55	1.87%	0.5%	0.56	-1.8%	–	–	
S87.C	1.88%	0.54	1.87%	0.5%	0.56	-3.7%	0.67	-13.0%	

Table 3.23: Validation for eigenvector values in scenario S2.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)			
	Eigenv. 1	Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2
S85.C	0.25	0.03	0.25	0.0%	0.05	-8.0%	–	–	–	–
S84.C	0.21	0.06	0.21	0.0%	0.07	-4.8%	0.24	-13.50%	-0.08	220.74%
S82.C	0.15	0.10	0.15	0.0%	0.10	0.0%	–	–	–	–
S81.C	0.15	0.10	0.15	0.0%	0.10	0.0%	0.17	-16.62%	-0.12	220.74%
S83.C	0.15	0.10	0.15	0.0%	0.10	0.0%	0.17	-16.62%	-0.12	220.74%
S80.C	0.14	0.11	0.14	0.0%	0.10	7.1%	0.16	-16.62%	-0.12	220.74%
S66.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	–	–	–	–
S64.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	0.15	-16.62%	0.01	220.74%
S63.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	–	–	–	–
S65.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	0.15	-16.62%	0.01	220.74%
S160.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	–	–	–	–
S62.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	0.15	-18.37%	0.01	220.74%
S60.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	0.15	-18.37%	0.01	220.74%
S61.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	–	–	–	–
S61s.C	0.13	-0.05	0.13	0.0%	-0.01	-30.8%	0.16	-20.94%	0.01	216.50%
S78.C	0.12	0.12	0.12	0.0%	0.11	8.3%	0.14	-18.70%	-0.13	216.50%
S79.C	0.12	0.12	0.12	0.0%	0.11	8.3%	0.14	-18.50%	-0.13	216.50%
S77.C	0.11	0.12	0.12	-9.1%	0.11	8.3%	–	–	–	–
S57.C	0.10	-0.04	0.10	0.0%	0.11	-150.0%	0.11	-14.00%	-0.13	221.00%
S91.C	0.10	0.13	0.10	0.0%	0.11	20.0%	0.11	-11.45%	-0.13	221.00%
S95.C	0.10	0.13	0.10	0.0%	0.11	20.0%	0.11	-11.45%	-0.13	221.00%
S92.C	0.10	0.13	0.10	0.0%	0.11	20.0%	0.11	-11.45%	-0.13	221.00%
S93.C	0.10	0.13	0.10	0.0%	0.11	20.0%	0.11	-11.45%	-0.13	221.00%

Table 3.24: Validation for voltage variation and average normalised covariance in scenario S6.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)		
	Voltage variation	Average norm. $\Sigma_{i,j}$	Voltage variation	Diff. voltage variation	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$	Average norm. $\Sigma_{i,j}$	Diff. average norm. $\Sigma_{i,j}$	
S66.C	-12.3%	0.99	-8.39%	46.9%	0.99	0.0%	–	–	
S65.C	-11.5%	0.94	-7.84%	47.2%	0.94	0.0%	0.99	-5.0%	
S64.C	-10.5%	0.88	-7.12%	47.7%	0.88	0.0%	0.87	1.0%	
S63.C	-9.7%	0.83	-6.52%	48.3%	0.83	0.0%	–	–	
S62.C	-9.2%	0.8	-6.22%	48.6%	0.80	0.0%	0.87	-7.0%	
S160.C	-8.6%	0.77	-5.80%	49.1%	0.77	0.0%	–	–	
S61.C	-8.6%	0.77	-5.80%	49.1%	0.77	0.0%	0.87	-10.0%	
S61s.C	-8.6%	0.77	-5.80%	49.1%	0.77	0.0%	–	–	
S60.C	-8.6%	0.77	-5.80%	49.1%	0.77	0.0%	–	–	
S57.C	-6.5%	0.64	-4.39%	49.3%	0.64	0.0%	0.87	-23.0%	
S85.C	-6.5%	0.61	-4.43%	46.2%	0.62	-1.6%	–	–	
S55.C	-5.6%	0.58	-3.72%	49.4%	0.58	0.0%	0.78	-20.0%	
S54.C	-5.6%	0.58	-3.72%	49.4%	0.58	0.0%	0.78	-20.0%	
S56.C	-5.6%	0.58	-3.72%	49.4%	0.58	0.0%	0.78	-20.0%	
S53.C	-5.2%	0.56	-3.48%	49.4%	0.56	0.0%	0.75	-19.0%	
S52.C	-4.6%	0.53	-3.10%	49.5%	0.53	0.0%	–	–	
S84.C	-5.2%	0.54	-3.56%	45.6%	0.55	-1.9%	0.74	-20.0%	

prior data on the system topology. These metrics are solely based on data measurements. One of the most interesting features of these metrics is their applicability to both radial and meshed systems. As demonstrated in section 3.4.6, these metrics can identify whether the distribution system, under the same perturbation, is radial or meshed based on the highlighted path values in M^P and M^Q.

Table 3.25: Validation for eigenvector values in scenario S6.

Node	1-minute measurements		10-minutes measurements (full)				10-minutes measurements (partial)			
	Eigenv. 1	Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2	Eigenv. 1	Error Eigenv. 1	Eigenv. 2	Error Eigenv. 2
S66.C	0.29	0.08	0.29	0.0%	0.08	0.0%	–	–	–	–
S65.C	0.27	0.07	0.27	0.0%	0.08	-3.7%	0.31	-13.50%	-0.10	223.87%
S64.C	0.25	0.06	0.24	4.0%	0.07	-4.2%	0.26	-16.62%	-0.07	222.65%
S63.C	0.23	0.05	0.22	4.3%	0.06	-4.5%	–	–	–	–
S62.C	0.21	0.05	0.21	0.0%	0.06	-4.8%	0.24	-16.62%	-0.07	222.65%
S160.C	0.20	0.05	0.20	0.0%	0.05	0.0%	–	–	–	–
S61.C	0.20	0.05	0.20	0.0%	0.05	0.0%	0.23	-16.62%	-0.06	222.65%
S61s.C	0.20	0.05	0.20	0.0%	0.05	0.0%	–	–	–	–
S60.C	0.20	0.05	0.20	0.0%	0.05	0.0%	–	–	–	–
S57.C	0.15	0.03	0.15	0.0%	0.04	-6.7%	0.17	-16.62%	-0.05	222.65%
S85.C	0.14	-0.07	0.14	0.0%	-0.05	-14.3%	–	–	–	–
S55.C	0.13	0.03	0.13	0.0%	0.03	0.0%	0.15	-18.37%	-0.04	218.73%
S54.C	0.13	0.03	0.13	0.0%	0.03	0.0%	0.15	-18.37%	-0.04	218.73%
S56.C	0.13	0.03	0.13	0.0%	0.03	0.0%	0.15	-18.37%	-0.04	218.73%
S53.C	0.12	0.03	0.12	0.0%	0.03	0.0%	0.15	-20.94%	-0.04	218.73%
S52.C	0.11	0.02	0.11	0.0%	0.03	-9.1%	–	–	–	–
S84.C	0.11	-0.07	0.11	0.0%	-0.06	-9.1%	0.13	-18.50%	0.07	215.70%

Similarly, as shown in section 3.5.3, voltage covariance can serve as an indicator of how far voltage variation is propagated across the system, and it can demonstrate the impact based on the system topology, regardless of whether it is radial or meshed. This is particularly useful when attempting to understand power distribution across a network with unknown topology or without previous data on how measurements are interconnected. However, these metrics cannot precisely identify the specific type of devices used in the system for voltage level control or power distribution, such as OLTC or capacitor banks. Nevertheless, if there are no measurements associated with these devices, the metrics can still reflect their presence in the system, as they will affect the magnitude of the metrics. Once again, this is highly valuable when dealing with an unknown or partially unknown distribution system that needs to be described based on the actual system conditions, rather than relying on general approximations or assumptions based on outdated data.

One potential application for these metrics is the development of real-time system models for prediction and control, especially when there is no available data from the system or when a purely data-driven model is desired for the entire system or a specific portion of it. The main advantage of these metrics is that they rely solely on measured data and are easy to implement. As demonstrated, even with power and voltage magnitudes alone, significant information about the system topology can be deduced, even in unbalanced conditions. A similar approach can be applied using other types of measurements, such as phasorial data from PMUs, which can further enhance the understanding of the system. In this

case, however, the metrics were developed using the most common type of data available in distribution systems worldwide. These metrics can be integrated into any system identification methodology to develop models based on regressions or non-linear descriptions of the system, as illustrated in the next chapter.

Although this data-based description provides a good understanding of real-time events, it can be influenced by various factors. The main limitation of the metrics is the number of available measurements from the system. As shown in section 3.5.4, as the number of measurements decreases, it becomes challenging to differentiate between different sections or portions of the system, thereby reducing the understanding of each developed path in the system. Therefore, these metrics require measurements to be installed in key parts of the system that need to be described. Additionally, these metrics are designed to evaluate critical system conditions, as the system response is reflected in the quasi-dynamics to be identified during significant perturbations. Consequently, it is required to evaluate the descriptions of events in the distribution system that do not experience high perturbations, which is part of the work done in next chapter. Furthermore, if there are multiple significant perturbations, the system tends to approximate the behaviour by combining both perturbations into a single description, requiring additional information to distinguish the contribution of each perturbation. In this thesis, the most extreme case was assumed (no previous description of the distribution system), and this can be significantly improved when historical data is available, as the user can deduce information regarding the number of loads, their locations, the nature of the perturbation, etc.

3.8 Discussion

The proposed metrics serve the purpose of characterising the spatial-temporal variations in voltage within a distribution system using limited measurements. The ultimate goal is to identify the nodes that are most crucial for observation and influential for control. These novel metrics, denoted as M^P and M^Q, are computed based on the available time series measurements of nodal voltages and power injections. The voltage covariance matrix, on the other hand, is derived solely from the time series voltage measurements. The objective of these metrics is to extract the most relevant characteristics of the system, such as the nodes that are most affected by a perturbation or control action, as well as the nodes that are most critical to monitor.

In terms of model development, utilising M^P and M^Q enables the identification of the most effective nodes for locating control actions, while nodes that exhibit

large covariance values can be identified as electrically close. This approach potentially allows the identification of the most critical nodes for control (which nodal power injections should be included in the control input vector, denoted as u) and observation (which nodal voltages should be included in the measurement vector, denoted as y). Consequently, it reduces the complexity of the model compared to a comprehensive whole-system model that aims to capture every node and line. For instance, nodes with controllable power injections, such as those from inverter-based devices, and with high values of M^P/M^Q should be considered for inclusion in the input vector u, while nodes with the highest average normalised covariance values should be considered for inclusion in the measurement vector y.

The objective of this analysis is to ascertain the maximum amount of information that can be extracted from the available power and voltage measurements in the distribution system. These metrics can serve as inputs for modelling since they provide spatial and temporal descriptions of the system's quasi-dynamics. The richness of the model can be enhanced when additional information about the system is available. For instance, if the topology of the distribution system, or at least the connectivity of nodes, is known, it becomes possible to reconstruct the voltage magnitude at critical nodes. Unfortunately, detailed and up-to-date information about the system is sometimes lacking.

The analysis of voltage magnitudes yields valuable insights due to the unbalanced electromagnetic compensation across all three phases. The magnitude of the Fisher z-transformation offers tentative information about the location of perturbations by analysing M^P and M^Q. This approach provides a deeper understanding of power distribution and the impact of voltage within the circuit. Patterns have been identified across various types of perturbations, asset configurations, and topology arrangements, making it applicable for both balanced and unbalanced analyses.

Furthermore, the power transmitted within the distribution grid is constantly changing, leading to variations in the impedance values. The conventional model approach relies on static knowledge of the distribution system with fixed parameters. However, assumptions such as transformers operating at rated values may not be realistic. Therefore, in order to facilitate the increased integration of renewable energy sources, a more accurate model that captures the dynamic nature of the system and captures the most relevant quasi-dynamics is required.

One of the significant advantages of the previously presented analysis is the utilisation of available measurements within the system. However, it is important to acknowledge the limitations associated with the number of measurable points

and the magnitude of perturbations, which are constrained by the information entropy limit [236]. Nevertheless, it has been concluded that having complete knowledge of the entire system, including voltage sensitivities throughout, is not necessary. It is sufficient to analyse the system's components in proximity to the perturbation point, as indicated by the electrical distance quantified through the average normalised covariance of measured voltages and patterns in power consumption/generation. This allows for the establishment of a criterion to cluster the system based on their proximity and group together nodes that are affected by a specific perturbation. This approach ensures accuracy without imposing limitations through localised models. The criteria for proximity will be flexible and adjusted according to the actual conditions of the system.

A Perturbation-Compensation (Actuator/Observation) approach for building voltage control models could be proposed using the information within the values of M^P and M^Q for the actuation vector, while values such as the average normalised covariance of the measured nodes can represented spatial characteristics of the system. Additional exogenous variables such as solar irradiance (in case of photovoltaic injections) and consumed power can be used to enrich the model approach. The output could be the prediction of voltages/ desired voltage to be controlled, which can be related in the time-series. After analysing the data and finding the linearity of the event under study, there are several ways to relate both inputs and outputs, such as using statistical analysis of time series model approach based on linear regression such as ARMAXs, or using Koopman-operator based approaches, which relates the data to obtain State-Space representations. This also can be highly improved by adding a layer of Non-linear PCA. Additionally, there are other approaches based on non-linear regression, such as using computational intelligence to model it, e.g., through ANN. NARMAXs can also be considered by adding Non-linear PCA to obtain the intrinsic coordinates of the system and ease the construction of the model.

In the next chapter, a methodology based on statistics analysis of inputs is presented to develop models using these approaches and validate its stability and performance.

3.9 Conclusions

This chapter focuses on utilising measured data to obtain insights into the controllability and observability of voltages in a distribution system. Specifically, the aim is to identify which nodal voltages are most affected by power injections or perturbations, as well as which voltages are electrically close and can be considered

similar in terms of measurements. The problem statement, which includes time-series data of electrical variables and exogenous agents impacting this relationship, is introduced in Sections 3.1 and 3.2. The generation of the data set used for analysis and the incorporation of uncertainties related to location and weather variables are explained in detail in Section 3.3. Notably, a probabilistic perspective is applied to assess voltage variations. It is important to consider that undesired voltage levels may not always correspond to the most probable case, which must be taken into account when constructing the model.

Section 3.4 examined the impact of power injections on the system and emphasises the significance of active and reactive power in potential control actions, taking into account the system's topology. New metrics, namely M^P and M^Q, are proposed to identify and quantify voltage perturbations at nodes resulting from power injections or consumption. These metrics require measurements of nodal voltages and injected line powers for their calculation. Section 3.5 explored alternative concepts of electrical distance to assess the level of connectivity between measurable nodes in the distribution system. A covariance matrix of nodal voltages, which is normalised and averaged, is proposed as a useful proxy measure for electrical distance. This matrix enables the identification of "electrically close" nodes, implying that not all nodes need to be observed to estimate the system voltages accurately. The metrics developed in this study were evaluated using different components and topological configurations, demonstrating their effectiveness in describing the system under various operational conditions. Section 3.6 presented a validation of the metrics, comparing them to a high-resolution reference case. The results showed that even with partial measurements, the metrics consistently ranked the relevance of components during the analysed scenario. Section 3.7 discussed the opportunities and limitations of implementing these metrics in real applications, highlighting the importance of measurement granularity and the detection of relevant system elements to be modelled. Despite these challenges, the metrics proved to be valuable for developing real-time models in situations where limited information, such as topology, is available and accurate predictions of high-perturbation impacts are desired.

Sections 3.8 and 3.9 provided a relevant discussion on the potential use of these metrics and presented conclusions based on the obtained results. The data-driven approach for analyzing voltage variations and power transmission between nodes was emphasized. The proposed approach was evaluated for different types of power and voltage fluctuations, and the simulation results consistently supported the effectiveness of the procedure. Furthermore, the results demonstrated how to maximize the utilization of available data to describe the distribution system

in real-time, identifying key data requirements for model construction, such as maximum voltage variation during defined perturbations.

While some of the proposed scenarios with limited measured data highlighted the need for additional data inputs to develop accurate models, the data-driven approach presented remarkable potential for reducing model complexity and capturing the necessary quasi-dynamics. By defining a criterion for system clustering and integrating multiple input/output regression methods, control models can be established. The study aimed to provide an alternative to complex techniques used in system reduction, such as moment matching or Hankel matrix, which may require advanced control theory knowledge and can be challenging to understand for distribution system operators. The primary motivation was to enable researchers to establish a reference for characterising key parameters relevant to constructing time-series control models while preserving the physical interpretation of the distribution system.

Chapter 4

Time-series modelling application in distribution systems

4.1 Introduction

In Chapter 3, new metrics were introduced to describe the distribution system. The subsequent task is to develop models based on this description. The implementation of ANM systems has facilitated the utilisation of real-time data for monitoring and control purposes [69]. However, ANM systems rely on physically-based models, which pose challenges in terms of maintaining up-to-date models. Moreover, these models make forecasts based on partial measurements available in the system. In general, keeping the system model updated and installing measurements throughout the system can be costly. Consequently, constructing models for ANM systems under such conditions presents a challenge for distribution system operators [8].

The information provided by traditional metrics, such as electric parameters, and other measurable data used to describe the distribution system, is subject to temporal changes. Consequently, the application of time-series analysis tools becomes essential. Time-series analysis allows the extraction of meaningful statistics and relevant system characteristics in the face of constant changes, as observed in electric networks [170, 171]. These extracted features can then be used to identify control models. With the significant increase in DERs participating in distribution systems, their integration has become a reality in distribution system operations. The uncertainties and high variability associated with renewable energy sources have intensified the interest in analysing the statistical behaviour of time-series data obtained from available measurements, including the conventional Probability density function (PDF) associated with load consumption [161].

Using time-series measurements of system variables for constructing control models has been extensively used and validated in industrial applications outside the power domain [15–18]. While time-series analysis has also found utility in power distribution network modelling and analysis, particularly in support of power flow analysis considering the presence of renewables [161], topology detection [98, 172, 173], and reactive power control [174], these analyses do not encompass the identification of the "broad dynamics" of a distribution system based on limited available time-series data, considering the effects and time-evolution of uncontrollable exogenous variables. As mentioned in Section 1.3.1, these "dynamics" differ subtly from the traditional notion of power system dynamics, as they encompass not only the dynamics of traditional assets but also capture the dynamic effects of loads and renewable inputs on system variables, such as voltage. They reflect the long-term interaction of all these assets, which is often referred to as "quasi-dynamics". For the purposes of this thesis, unless stated otherwise, the term "dynamics" or "quasi-dynamics" in this chapter corresponds to this definition.

The analysis required to produce these models should be divided into two parts: the first by extracting relevant variables using statistical analysis; the second by generating the model based on regression techniques. It is common to define the selected explanatory variables in the modelling and their relevant time lags [238–240]. Research done in transmission systems intuitively integrates this concept [240, 241], and other applications considering time lags are used to predict electricity price, which is a strong field studied in economics [242, 243]. Developing causal analysis in time-series modelling is a challenging task [171, 244–246], for which is not required a deep “causation” study in the context of power systems. The main question for this task should be, is it possible to predict a variable due to its interaction with another (measurable) explanatory variable, in different time lags?

Regarding using regression techniques to predict variables, most of applications are focused on load forecasting using statistical models [142, 145–147, 245, 247–252]. These regressions are highly impacted with the load type or size. There was also an attempt of directly model voltage PDF based on knowledge of topology according to [236] or evaluating different variables such as rated power in transformers and lines[253]. Additionally, computational learning-based methods are also used [239, 254–259] to fit a regression that represent the data. However, these methods do not provide an explanation for the underlying nature that generates the distribution function. While these methods are directly applied to predict the desired parameter, they offer a broad approximation of the variable for a portion of the distribution system. There is a need to develop representations that

also consider the spatial characteristics of the variables. Hence, a comprehensive analysis of all variables is necessary to determine which variables can effectively explain the quasi-dynamics of the desired parameter. Acknowledging this reality, an alternative question and perspective addressed in this thesis is whether it is possible to develop purely data-based models using only measurable variables, including both endogenous and exogenous variables, that *are* available in the system?

Proposed variables are presented in Chapter 3 and used as a start point to produce the desired models. It is proposed a methodology to select the relevant data to be considered in the modelling approach. The objective is to develop a model that predicts voltage and can integrate exogenous variables and control inputs. Since several variables could explain same of the quasi-dynamics, they should be removed and grouped to simplify the model (collinearity problem). This revision includes also checking relevant properties, such as stationary, heteroscedasticity, and normality. With the data "cleaned", relevant lags are revised using the Granger-causality concept. With relevant variables and lags selected, data regression models are performed using different techniques, some of them based on classic representations such as ARX, ARMAX and others based on Koopman operator representations, e.g., DMD. A comparison of performance using statistical tools explains the model's validity. Finally, distinct types of model configuration are evaluated, i.e., Multiple-Input and Single-Output (MISO), Multiple-Input and Multiple-Output (MIMO). Using the information available, it is considered an unbalanced network with an arbitrary penetration level from renewable power sources and assumed nothing about its topology and parameters. It is investigated the efficacy and validity of the proposed methodology via case study simulations on a 123-bus test network.

This chapter presents an approach that relies solely on measured data to construct a reduced-order representation of the system for voltage control in an unbalanced distribution system. The main contribution of this chapter is the introduction of a methodology for developing data-driven models for distribution system applications, based on statistical analysis of measurable data and exogenous data. These models aim to reconstruct the desired quasi-dynamics and predict/control certain system variables, such as voltage levels. Typically, the development of such models involves working with partial knowledge of the system (e.g., presuming system topology). This often requires estimating the missing variables or using non-linear regression techniques, such as artificial intelligence, to build a representation that aligns with the available data. In contrast, this chapter introduces a statistically supported regression approach that can linearly predict voltage based

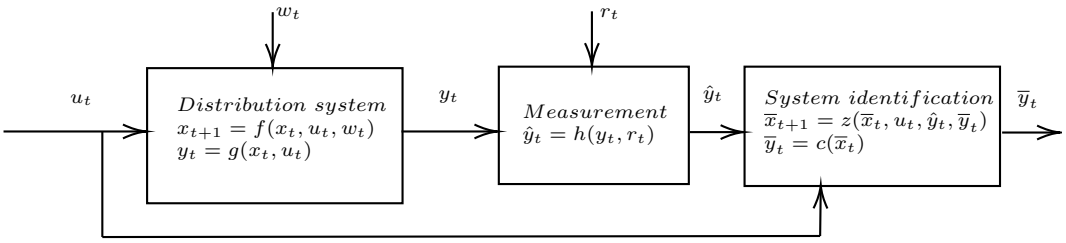


Figure 4.1: Description of the modelling approach scheme

on the measurable data. The key contributions of this chapter towards achieving this methodology are as follows:

1. It is contrasted response analysis of static and time-variant responses to define relevant lags, using cross-correlation analysis and contrasted with Granger-causality analysis;

2. It is proposed a *data-driven* approach to get a reduced-order linear representation of the distribution systems that consider exogenous variables to get one-step voltage ahead; and

3. It is verified some general assumptions in the statistical approach and presented a methodology to improve the response based on the analysis of time-series data.

4.2 Problem statement

A general representation of the problem is presented in Figure 4.1. It is considered a general distribution network wherein the voltage–power quasi-dynamics are assumed to be governed by DAEs presented in equation (3.1).

These DAEs capture the physics of the electrical network and the time-varying, non-deterministic and dynamic actions of consumers and producers. For practical purposes, equation (3.1) that represent the distribution system will be presented in this chapter in its discrete-time representation; t denotes time, x is a vector of (internal) states, whose time evolution is described by equation (3.1a), u is a input vector that contains exogenous variables (and potential control inputs), which in this case are active/reactive power injections at nodes that can be measured and/or controlled, the irradiance, and the new metric proposed in Chapter 3, y is vector collecting the voltages at each node $i \in \{1, \ldots, N\}$, and w is a vector of uncertain variables affecting the state evolution.

The motivation for this thesis is to perform voltage predictions that can be used in control of the distribution network without knowledge about its quasi-

dynamics introduced in equation (3.1) are unknown. Moreover, the system should comprise several unmeasured states, i.e., only a subset of the nodal voltages is measured, and other subset of nodal power injections is available for measuring (and control) purposes. All measurements are stored in vector $\hat{y}$, which include a subset of measured values of y, and it is affected by the vector of uncertainties r.

The aim of this chapter is to perform a *system identification* process, which is required to construct a simple yet sufficiently accurate model of the power–voltage quasi-dynamics at the timescale of interest for voltage regulation. Therefore, the system identification will produce a model according to the DAEs:

$$\bar{x}_{t+1} = f(\bar{x}_t, u_t, \hat{y}_t, \bar{y}_t) \tag{4.1a}$$

$$\bar{y}_t = c(\bar{x}_t) \tag{4.1b}$$

In equation (4.1), $t+1$ denotes time one-step ahead, $\overline{x}$ is a vector of (internal) states, whose time evolution is described by equation (4.1a) and represent the reduced order model of the original system in equation (3.1). $\overline{y}$ is a vector collecting the predicted voltages at measured nodes. Further explanations of distribution systems to be identified and measurement conditions are presented in Sections 3.2 and 3.4.4.

In this chapter, the inputs and outputs used to build the models are discussed in more detail in the following sections. The inputs encompass all available historical data used to describe the studied quasi-dynamics. Specifically, these inputs include the magnitude of voltage at various nodes, power consumption by the load, power injected into the system by PV units, and solar irradiance. These inputs complement the information provided by the newly proposed metrics M^P and M^Q, as well as the average normalised covariance. The output of the model is the predicted voltage. For the purpose of this chapter, all the data used are synthetic/generated time-series data, following the procedure presented in Section 3.3. However, in real applications, historical data from actual or measured time-series can be utilised instead. More detailed information about the data used to generate the model is presented in Section 4.3, while the proposed methodology and its inputs and outputs are explained in Section 4.4.

4.3 Checking of data input in the modelling approach

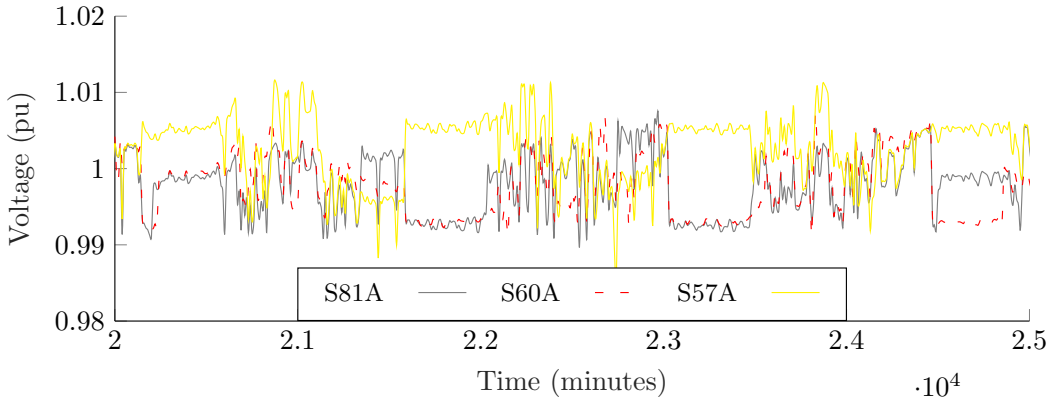
Before proceeding with the system identification process, it is required to examine the characteristics of the data. Time-series simulations were performed using OpenDSS, and the results were processed using MATLAB following the method-

ology described in Section 3.3. The IEEE 123-node system, shown in Figure 3.3, was used as the reference system. The locations of power injections and the assignment of generation/consumption profiles followed the procedure outlined in that section. For this chapter, a fixed renewable penetration level of 30% of the total rated power of loads was established, resulting in the installation of 17 randomly assigned PV generation units throughout the system. For this case, the OLTCs are connected and operating, while the capacitor banks are not connected nor there is any meshed component in topology of the system. It is noted that the total penetration level is calculated based on the total rated power installed across the distribution system. While a specific PV unit may inject more power than a particular load, the total installed generation capacity in the system represents a percentage of the total rated power of all loads. In this case, the 17 installed units collectively contribute to 30% of the total power consumed by all loads.

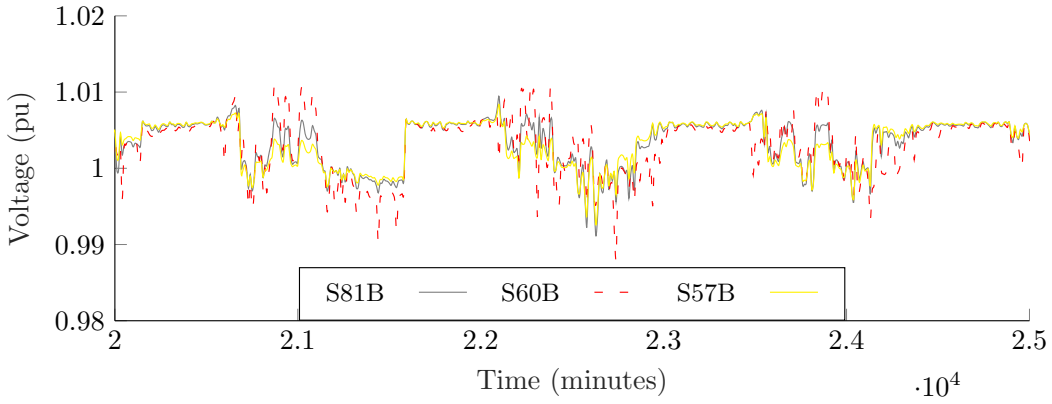
One of the PV units was located at node 85, phase C. The installed PV units are listed in Table 4.1. The irradiance data for the PV units is calculated assuming the network's location is Sheffield, using the CREST model introduced in the previous chapter. The time-series simulations were conducted for 1000 typical summer weekdays, and the representation of two arbitrary days is shown in Figures 4.2, 4.3, 4.4 and 4.5. Figures 4.2 and 4.3 illustrate the input data used for the modelling, providing insights into the nature of power, voltage, and exogenous variables (except for irradiance, which is not represented here). Figures 4.4 and 4.5 display the proposed metrics obtained from the time-series data. Similar to the previous chapter, the measurement sampling time in this case is set to 10 minutes.

Table 4.1: Assignation of PV units installed across the system

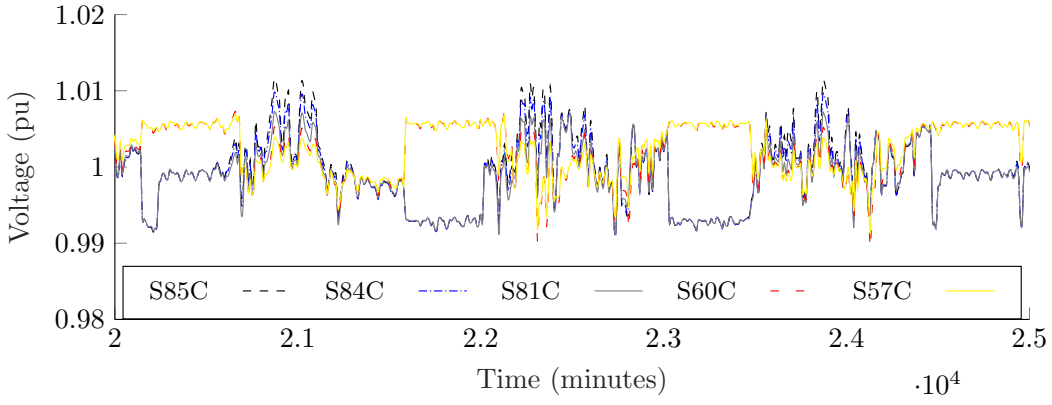
Name and location of PV unit	Size (kVA)
PV S85.C	87.77
PV S38.B	87.77
PV S49.A	52.61
PV S100.C	87.77
PV S4.C	61.89
PV S7.A	87.76
PV S80.B	45.72
PV S111.A	87.77
PV S76.A	50.80
PV S76.B	50.97
PV S53.A	81.93
PV S68.A	87.77
PV S65.A	5.68
PV S65.C	5.61
PV S42.A	87.77
PV S59.B	68.55
PV S99.B	87.77



(a) Voltage profiles of measured nodes, phase A



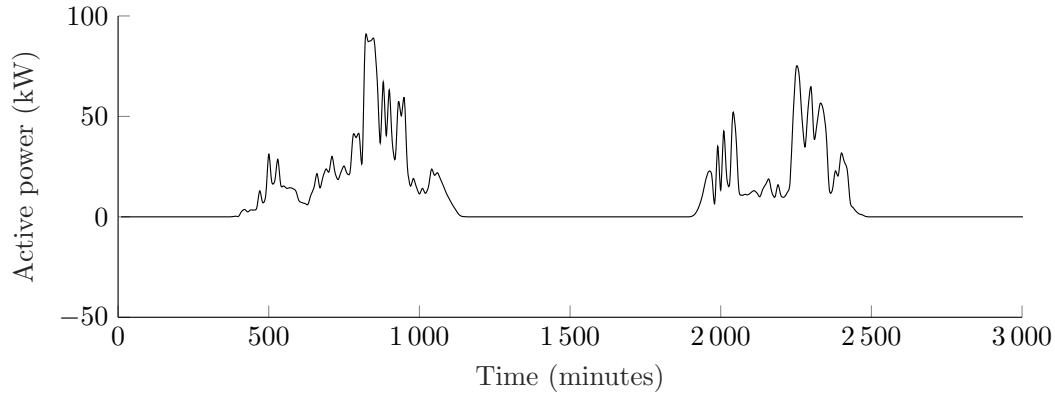
(b) Voltage profiles of measured nodes, phase B



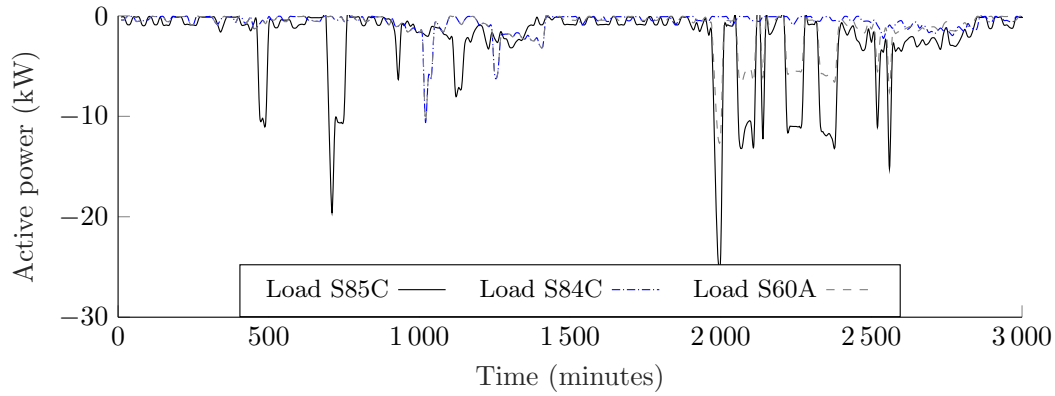
(c) Voltage profiles of measured nodes, phase C

Figure 4.2: Arbitrary selection of data voltage results after simulation with a penetration level of 30%

An exploration of critical measurable point is required to understand the capacity of the proposed approach. The measurement points selection is based on location in the system (end and middle of feeder), where they are normally located



(a) Active power profiles of measured (generation) at node 85, phase C

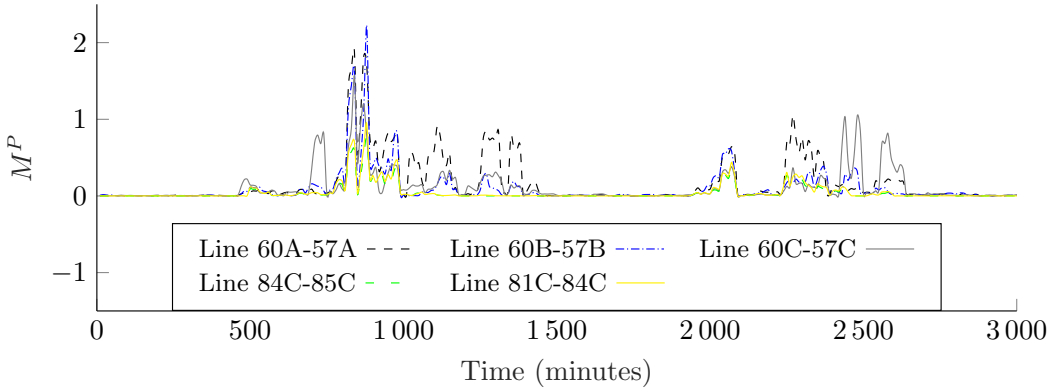


(b) Active power profiles of measured (consumption)

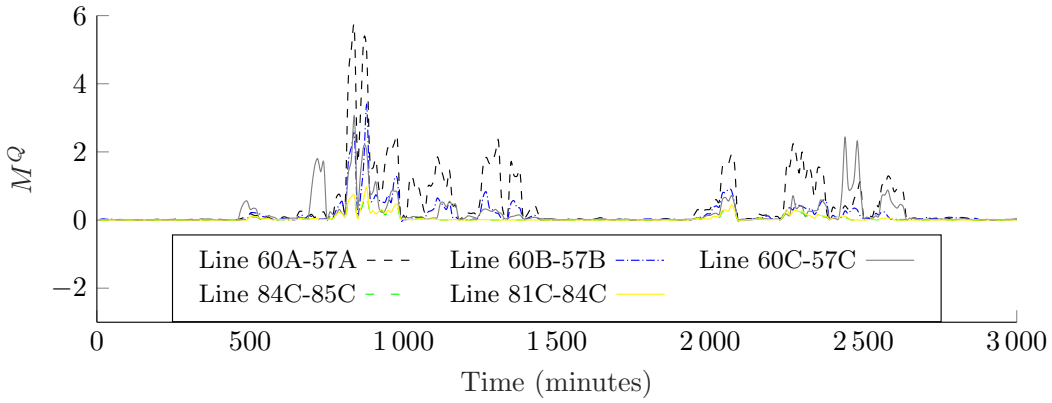
Figure 4.3: Arbitrary selection of data results of power consumed and generated from measured nodes after simulation with a penetration level of 30%

in real distribution systems. Responses over different nodes or lines in different phases will not change considerably the modelling methodology approach tested in this chapter. The optimal location of measurement points is not part of this thesis scope. Therefore, it is assumed that voltage measurements units are accessible only at nodes 57, 60, 81, 84 and 85. Equivalently, power measurements over lines 60-57, 81-84, and 85-84 are the only available measurements. The power measured in the system corresponds to load consumption at nodes 84 and 85, phase C, node 60, phase A, and the power injected for a photovoltaic generation unit at node 85, phase C. Additionally, the irradiance level was measured. The input vector u includes the time-series data of exogenous variables of consumed/injected power, the irradiance levels and the proposed new metrics (M^P,M^Q and the average normalised covariance). The measured vector $\hat{y}$ corresponds to time-series data of the voltages from vector y.

When creating time-series models, it is assumed stationarity of data [246].



(a) Profile obtained for metric M^P in different phases



(b) Profile obtained for metric M^Q in different phases

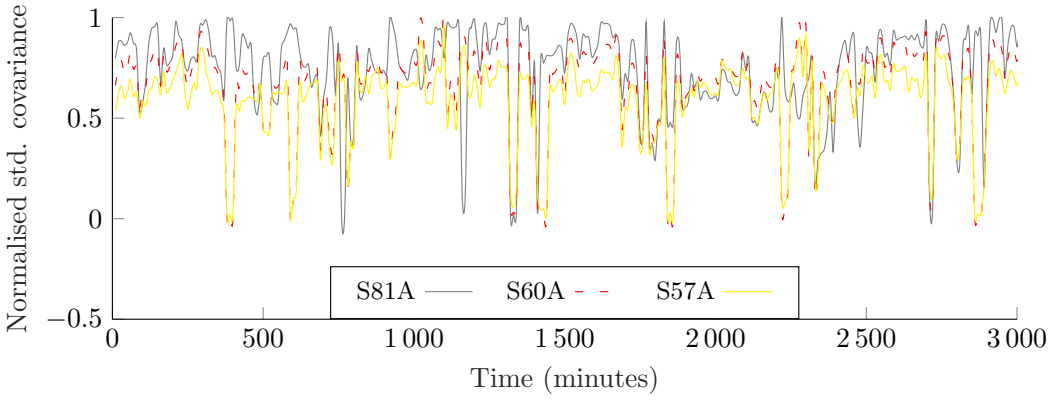
Figure 4.4: M^P and M^Q metrics obtained from arbitrary selection of data results

Thus, the data must present an autocorrelation structure and constant mean and variance. To understand how the suggested model approach should be constructed, a revision of this statement is required in this context.

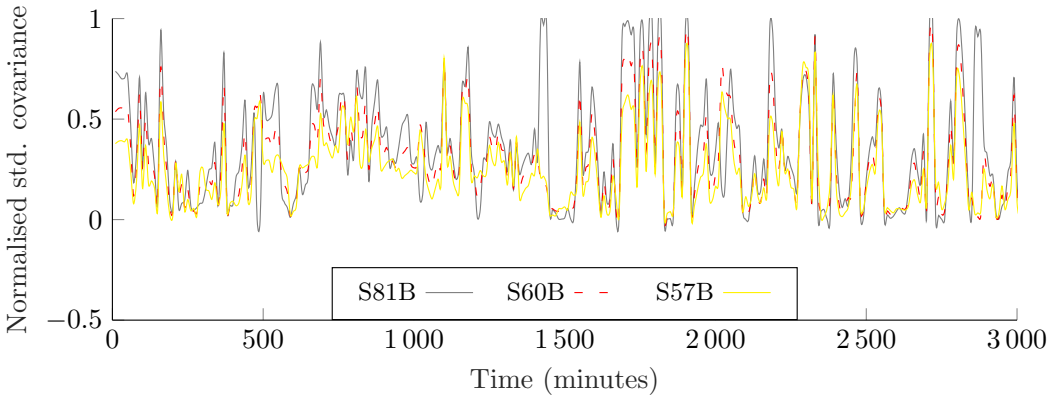
The analysis is done only in phase C as an illustration, however, the outcome is the same for the other two phases. Any time series can be described according to its trend-cycle component T_t, its seasonal component S_t and its remainder component R_t in the additive decomposition as shown in equation (4.2):

$$y_t = T_t + S_t + R_t \tag{4.2}$$

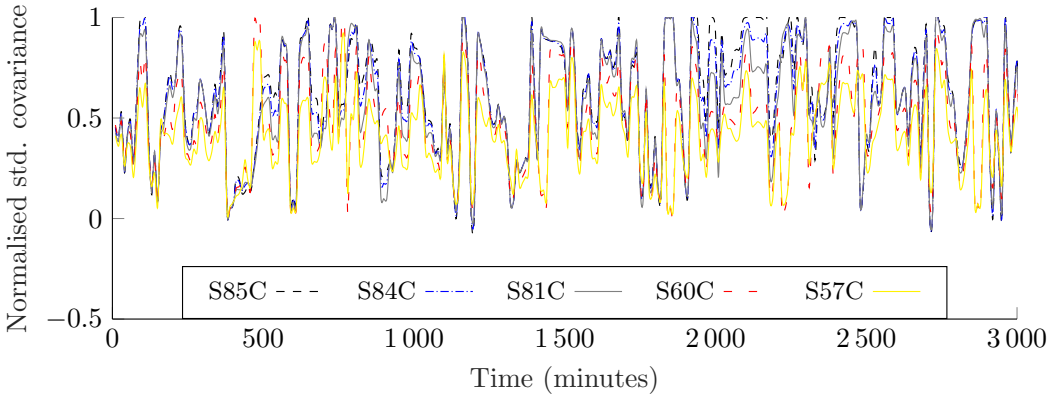
The original intention of dividing the data into weekdays/weekend and summer/winter cases was to mitigate the influence of seasonality patterns, which can complicate the construction of models. However, other patterns are still observed, as depicted in Figures 4.6 and 4.7. These figures illustrate the decomposition of a portion of the voltage data into its total components. Figure 4.6a displays the



(a) Profile obtained for Phase A



(b) Profile obtained for Phase B

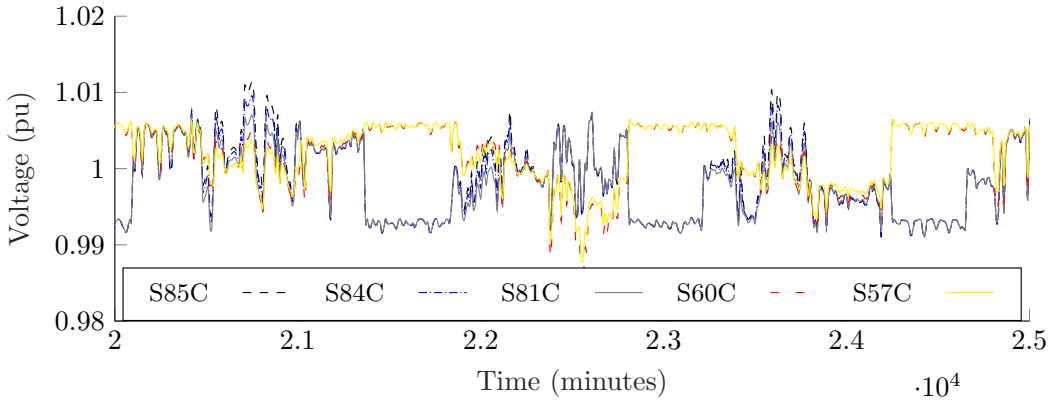


(c) Profile obtained for Phase C

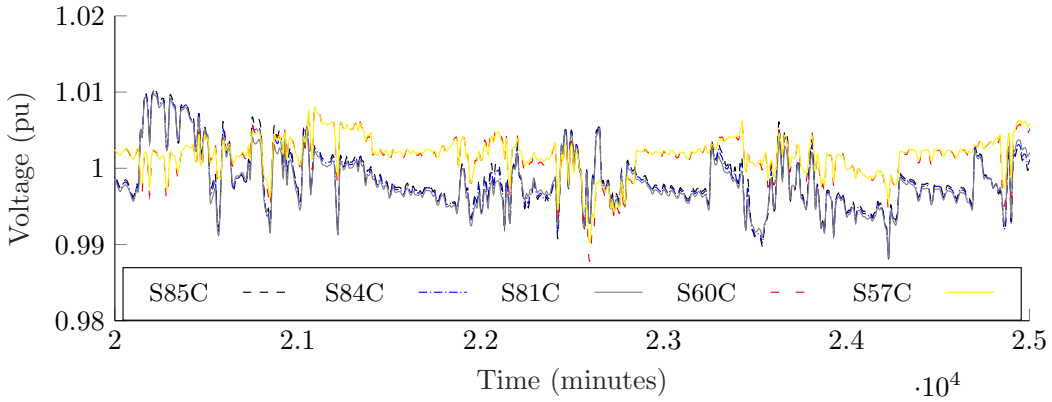
Figure 4.5: Average normalised standard metrics obtained from arbitrary selection of data results

total components, while Figures 4.6b,4.7a, and4.7b represent the individual components T_t, S_t, and R_t, respectively.

The presence of these components in the data can be attributed to certain pat-



(a) Measured voltage



(b) Trend component T_t

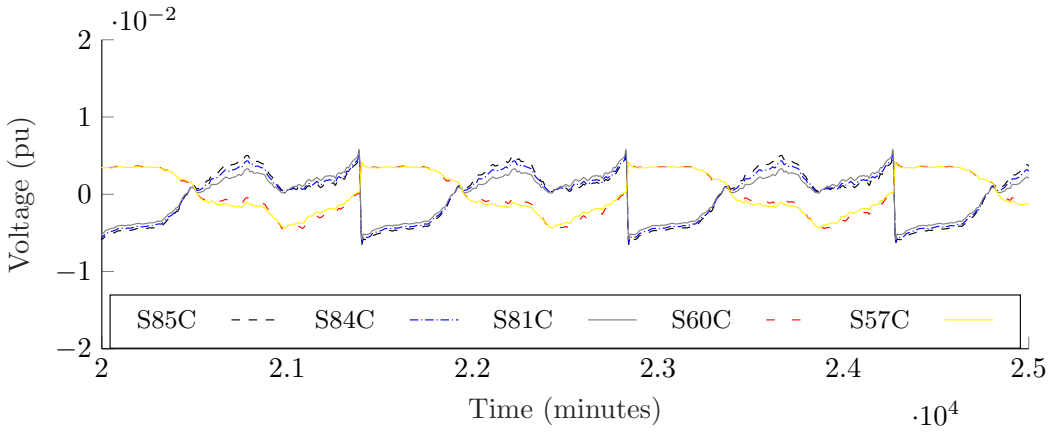
Figure 4.6: Voltage reference and first component (trend) of measured voltages from arbitrary selection of data results

terns that repeat within a 24-hour period. Although the voltages tend to remain within a range of 0.99 and 1.01, as shown in Figure 4.6b, other characteristics related to cyclic patterns in consumer consumption and solar irradiation profiles are captured by the seasonal component S_t, as depicted in Figure 4.7a. The remainder component shown in Figure 4.7b is associated with the stochastic nature of random variables representing consumer behaviour and renewable generation units.

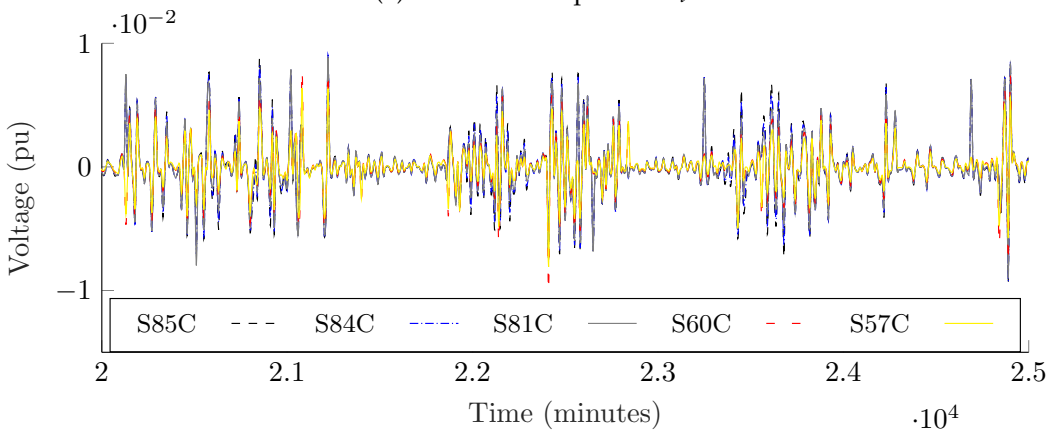
There are several approaches to deal with this. One of the simplest is data differencing [246], which means that a data value D at time t is differenced in order one, as indicated in the equation:

$$\Delta D_t^{(1)} = D_t - D_{t-1} \tag{4.3}$$

The data trend has been removed after differencing and is now stationary.



(a) Seasonal component S_t



(b) Remainder component R_t

Figure 4.7: Second and third components (seasonal and remainder) of measured voltages from arbitrary selection of data results

This eases the model creation since it will not have any AutoRegressive Integrated Moving Average (ARIMA) structure but a shape of a regression model in difference with ARMA errors. Therefore, it is required that all data used must be differenced in the same way as indicated in equation (4.3).

4.4 Proposed methodology for time-series data modelling

The proposed methodology consists of five steps, involving the revision, selection, organisation, and preparation of all measured data for use in a linear regression approach. Once the model is created, statistical assumptions are verified to determine its suitability for predicting and controlling voltage levels at each node. The procedure is summarised in Algorithm 4.1. This algorithm should be executed whenever there is a need to obtain a representation of critical nodes in the distri-

bution system, such as when measured or historical data indicate potential operational limit issues. For real-time applications, the algorithm can be re-evaluated at each sampling interval or according to a predefined schedule based on available historical data.

To build the model, the minimum amount of required data is determined by the production of the proposed metric introduced in Chapter 3. For this chapter, a window size of thirty minutes is used, which corresponds to three samples considering that data is sampled every ten minutes. However, it is important to note that this algorithm relies on having a sufficient amount of data to improve the performance of the model. To illustrate the application of the algorithm, it is assumed that the previously introduced measurements are critical nodes that are the only ones being monitored (and potentially controlled), and there is historical data spanning one thousand days.

For linear regression, it is desired that the estimators used as input follow a normal distribution since they help to obtain optimal response and produce results that can be easily analysed (e.g., the definition of a prediction interval), which is hard to obtain [260]. The most relevant assumption after building the model is that the obtained residual follows a white noise structure. Then, normality must be checked over residuals to validated a Gaussian distribution function of these random variables. Additionally, heteroscedasticity should be also checked, to confirm that is not present in the residuals, i.e., the variance is equal over the range of measured values.

Figure 4.8 illustrates the proposed approach for obtaining models, which encompasses steps 1-3 in Algorithm 4.1. Table 4.2 provides a summary of the inputs and outputs considered in the modelling process. The first step involves introducing the input data at time t, which includes control variables, exogenous variables, and the output voltage to be predicted. The next step is data pre-processing, where relevant measurements are selected after detrending or ensuring stationarity. A collinearity analysis is then performed to address issues of multicollinearity, which can increase variance unnecessarily due to the presence of redundant inputs [260]. Additionally, the relevant lags for better response are determined through cross-correlation analysis and Granger-causality analysis [246]. Once the data has been pre-processed, the next stage is to select data for training and validation purposes. In this case, the data is divided into 50% for training and the remaining 50% for validation. Finally, linear regression is performed using relevant techniques such as polynomial-based regression [102, 104] or Koopman-operator-based re-

Algorithm 4.1 Data-driven time-series modelling approach

Input: Historical data: V, P_K^D, P_L^G, *Irradiance*, M^P, M^Q, *Av.norm.cov*,
Output: Model representation (State-Space), Predicted V
initialization
Step 1 - Data revision and pre-processing
while *Data stationary==false* **do**
| Difference data according to (4.3)
end
Evaluate first attempt of linear regression
Check residuals properties (Autocorrelation, Heteroscedasticity, Normality)
if *Residual == white noise* **then**
| The model is completed, and it is fully explained statistically speaking
else
| Data requires processing
end
Step 2 - Data processing and selection
Selection of critical data to be modelled
Check balance of training/validation datasets
Check normality properties of data
if *Data == normal distribution* **then**
| Do nothing
else
| Transform to normal distribution using (4.63)
end
Check collinearity in used data
if *Data collinearity == true* **then**
| Remove redundant data
else
| Do nothing
end
Check relevant lags (model order) using cross-correlation analysis and Granger causality test
Step 3 - Creation of Linear time-invariant (LTI) model using revised data
Select I/O relationship (MISO, MIMO)
Apply regression method (Autoregressive-based, Koopman-based, etc)
Step 4 - Checking validity of assumptions
Check residuals properties (Autocorrelation, Heteroscedasticity, Normality)
if *Residual == white noise* **then**
| The model is completed, and it is fully explained statistically speaking
else
| More data/info is required to explain the quasi-dynamics/ increase horizon of prediction
end
Step 5 - Obtaining prediction interval for the time-series modelling
Calculating predicted voltages according to (4.4)

gression [136, 261], following the specified structure

$$\overline{y}_{t+1} = \overline{y}_t + \overline{\Delta y}_{t+1} \tag{4.4}$$

where the term $\overline{\Delta y}_{t+1}$ corresponds to the variation obtained from the LTI system model obtained in State-Space form

$$\overline{x}_{t+1} = A\overline{x}_t + B\Delta u_t + w_t \tag{4.5a}$$

$$\overline{\Delta y}_t = C\overline{x}_t + D\Delta u_t + e_t \tag{4.5b}$$

where w_t and e_t are assumed to be white noise for process and measurement. The matrices A, B, C, and D play specific roles in defining the behaviour of the system. The matrix A represents the quasi-dynamics of the system, describing how the state of the system evolves over time. The matrix B captures the effect of actuation, indicating how control inputs impact the system's dynamics. The matrix C defines the sensing strategy, determining which states are measured or observed. Finally, the matrix D represents the effect of actuation feed-through, accounting for any direct influence of control inputs on the output. When making predictions using the state estimate $\overline{x}_{t+1}$, it is assumed that the current information of the plant is required for accurate predictions. This assumption implies that the input cannot instantaneously affect the output. Consequently, in the plant model, it can be assumed $D = 0$, indicating that there is no direct feed-through of the control input to the output. This assumption simplifies the model and assumes that the output is solely dependent on the system's internal dynamics and the control inputs indirectly through the state variables.

The dimension of the model produced after regression varies depending on the number of inputs and outputs selected during the preprocessing stage. Also, the definition of key/relevant measurements will depend on the scenario and type of perturbation done into the system. The internal parameters of the model, obtained through regression, respond to the current measured situation of the system. As a result, the model is not constrained by the system's topology (radial or meshed) or by unmeasured devices. The effects of these factors on the system response are captured in the measurements, as discussed in the previous chapter. This facilitates the process of model creation without sacrificing generality.

In order to explore the capacity of the proposed approach, it is necessary to consider the selection of critical measurable points in the system. For this thesis, the measurement points are chosen based on their location in the system, which typically corresponds to the end and middle of feeders in real distribution systems. The methodology and approach tested in this chapter are not significantly affected

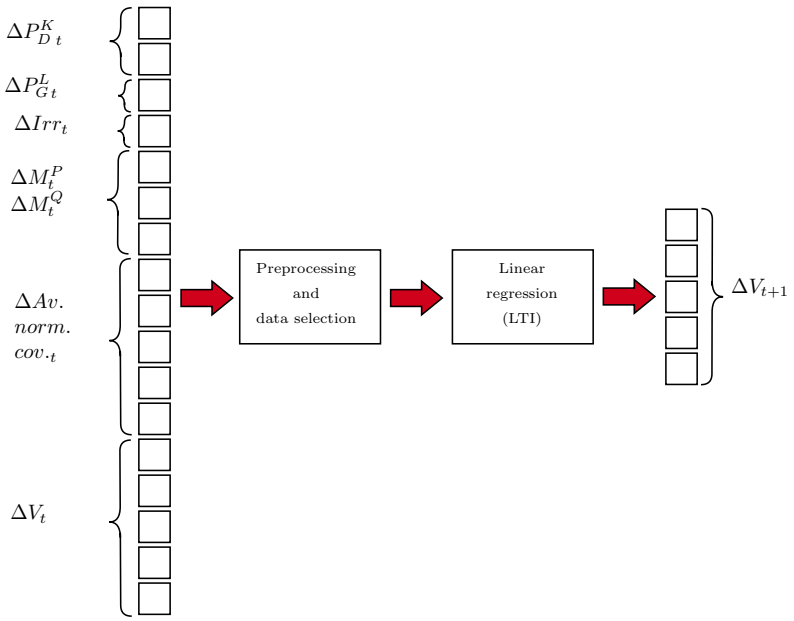


Figure 4.8: Proposed modelling approach for voltage prediction (Steps 1-3 of Algorithm 4.1)

by the specific nodes or lines selected for measurements. The optimal location of measurement points is beyond the scope of this thesis. In this case, it is assumed that voltage measurements are accessible at nodes 57, 60, 81, 84, and 85 in all available phases at each node. Similarly, power measurements are available on lines 60-57, 81-84, and 85-84. The power measurements in the system correspond to the load consumption at nodes 84 and 85 in phase C, the load consumption at node 60 in phase A, and the power injection from a photovoltaic generation unit at node 85 in phase C. Additionally, the irradiance level is also measured. The input vector, denoted as u, consists of the time-series data of the exogenous variables, including the consumed/injected power, irradiance levels, and the proposed new metrics (M^P, M^Q, and the average normalised covariance). On the other hand, the measured vector, denoted as $\hat{y}$, corresponds to the time-series data of the voltages from the vector y. The selected inputs and outputs for the modelling process are summarised in Table 4.2.

4.4.1 Data revision and pre-processing

The first step in this methodology corresponds to data revision and pre-processing. This step is summarised in the first step of the Algorithm 4.1.

As shown previously, data must be checked for stationarity. This could be checked visually, using Autocorrelation Function (ACF) and Partial Autocorrela-

Table 4.2: Selected inputs and outputs for linear regression

Inputs	Outputs
ΔPower consumed node 85 phase C	ΔPredicted voltage value node 85 phase C
ΔPower consumed node 84 phase C	ΔPredicted voltage value node 84 phase C
ΔPower consumed node 60 phase A	ΔPredicted voltage value node 81 phase A
ΔPower PV node 85 phase C	ΔPredicted voltage value node 81 phase B
ΔSolar irradiance	ΔPredicted voltage value node 81 phase C
$\Delta M^{P}_{60A-57A}$	ΔPredicted voltage value node 60 phase A
$\Delta M^{P}_{60B-57B}$	ΔPredicted voltage value node 60 phase B
$\Delta M^{P}_{60C-57C}$	ΔPredicted voltage value node 60 phase C
$\Delta M^{P}_{84C-85C}$	ΔPredicted voltage value node 57 phase A
$\Delta M^{P}_{84C-81C}$	ΔPredicted voltage value node 57 phase B
$\Delta M^{Q}_{60A-57A}$	ΔPredicted voltage value node 57 phase C
$\Delta M^{Q}_{60B-57B}$	
$\Delta M^{Q}_{60C-57C}$	
$\Delta M^{Q}_{84C-85C}$	
$\Delta M^{Q}_{84C-81C}$	
ΔAverage normalised covariance node 85 phase C	
ΔAverage normalised covariance node 84 phase C	
ΔAverage normalised covariance node 81 phase A	
ΔAverage normalised covariance node 81 phase B	
ΔAverage normalised covariance node 81 phase C	
ΔAverage normalised covariance node 60 phase A	
ΔAverage normalised covariance node 60 phase B	
ΔAverage normalised covariance node 60 phase C	
ΔAverage normalised covariance node 57 phase A	
ΔAverage normalised covariance node 57 phase B	
ΔAverage normalised covariance node 57 phase C	
ΔCurrent voltage value node 85 phase C	
ΔCurrent voltage value node 84 phase C	
ΔCurrent voltage value node 81 phase A	
ΔCurrent voltage value node 81 phase B	
ΔCurrent voltage value node 81 phase C	
ΔCurrent voltage value node 60 phase A	
ΔCurrent voltage value node 60 phase B	
ΔCurrent voltage value node 60 phase C	
ΔCurrent voltage value node 57 phase A	
ΔCurrent voltage value node 57 phase B	
ΔCurrent voltage value node 57 phase C	

tion Function (PACF). According to [262], the autocorrelation r_k for lag k between the univariate time series and stochastic process y_t and y_{t+k}, where $k = 0, \cdots, K$ is defined as

$$r_k = \frac{\frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y}_m)(y_{t+k} - \bar{y}_m)}{c_0}, \tag{4.6}$$

where $\bar{y}_m$ indicates the mean of y and c_0 is the sample variance of the time series. On the other hand, the PACF is defined as the autocorrelation between y_t and y_{t+k} with the linear dependence of y_t on y_{t+1} through y_{t+k-1} removed. That is

the same as calculating the autocorrelation between y_t and y_{t+k} that is not accounted for by lags one through $k-1$, inclusive:

$$\phi_{1,1} = \text{corr}(y_{t+1}, y_t), \text{ for } k = 1, \tag{4.7a}$$

$$\phi_{k,k} = \text{corr}(y_{t+k} - \widetilde{y}_{t+k}, y_t - \widetilde{y}_t), \text{ for } k \geq 2, \tag{4.7b}$$

where $\widetilde{y}_{t+k}$ and $\widetilde{y}_t$ are linear combinations of $\{y_{t+1}, y_{t+2}, \cdots, y_{t+k-1}\}$ that minimise the mean squared error of y_{t+k} and y_t, respectively. The theoretical stationary time series partial autocorrelation function can be calculated by using the Durbin–Levinson Algorithm [262].

Figures 4.9 and 4.10 show the corresponding voltage profiles and their corresponding ACF and PACF. More results for all phases are presented in Appendix J. As it is shown, there are high correlation levels for all lags in the ACF, a common behaviour in non-stationary systems. For higher-order lags that were not plotted, the data show how autocorrelation patterns periodically fluctuate and resemble a sinusoidal wave, and the significant partial auto-correlation in the lags were seasonal period restart, indicating seasonality.

4.4.1.1 Revision of non-stationarity of data

The analytical way to check the non-stationary condition corresponds to applying the Augmented Dickey-Fuller (ADF) test, which is a statistical significance test and indicates a failure to reject the null hypothesis that a unit root is present [171, 246, 260].

Consider the discrete-time stochastic process $(y_t, t = 1, 2, 3, \cdots)$, supposing that is represented by an autoregressive process of order p, $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t$, where ε_t is a serially uncorrelated, zero-mean stochastic process with constant variance σ^2. Assuming $y_0 = 0$ (for convenience), if $m = 1$ is a root of the characteristic equation of multiplicity 1 ($m^p - m^{p-1}a_1 - m^{p-2}a_2 - \cdots - a_p = 0$), then the stochastic process has a unit root.

The unit root is a property of a non-stationary time series that can lead to a wrong inference as a consequence of spurious regressions. Any unit root test evaluates if a time series variable possesses a unit root and it is non-stationary. The unit root test can be represented as shown in equation (4.2), where the basic concept of the unit root test is to determine whether the stochastic component consists of a unit root or not.

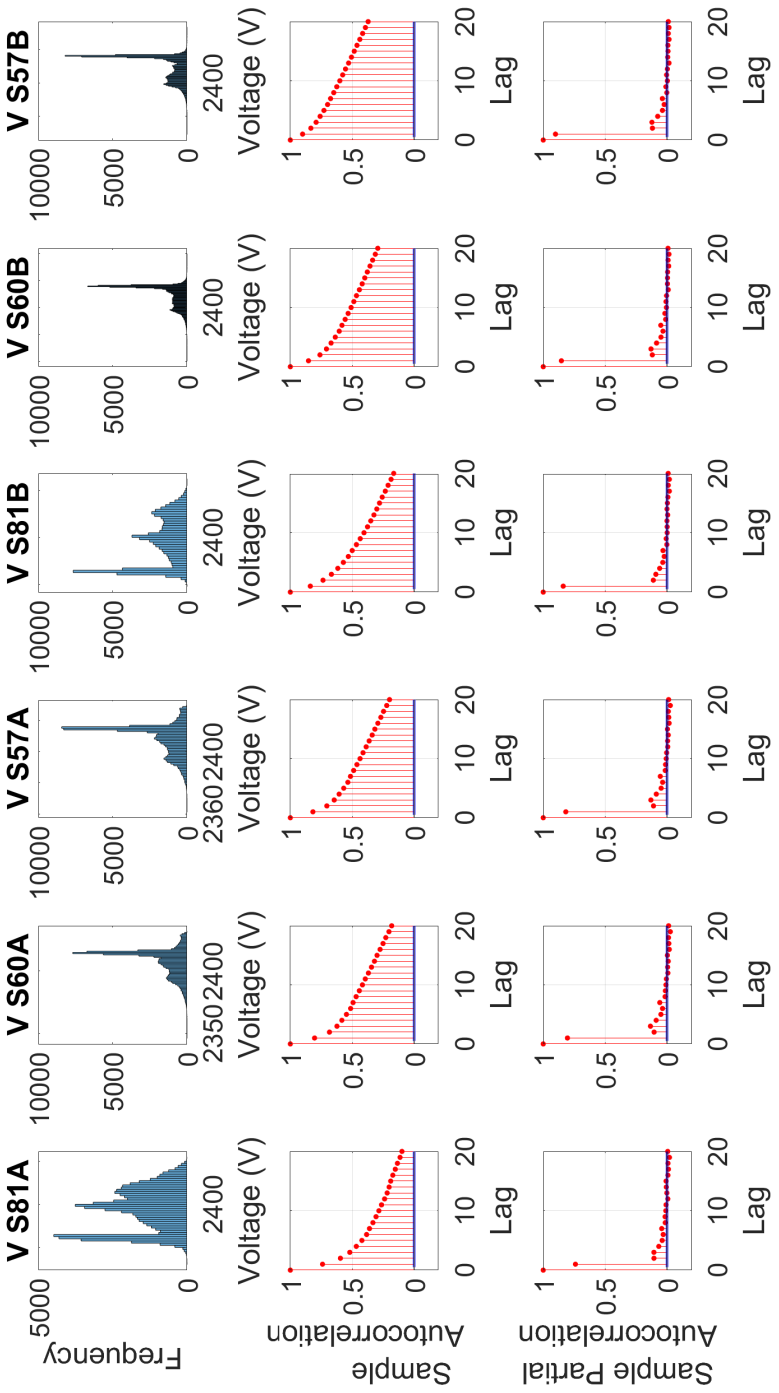


Figure 4.9: Distribution shape, ACF and PACF results of voltage distributions obtained from the reference case, showing non-Gaussian shapes and non-stationarity

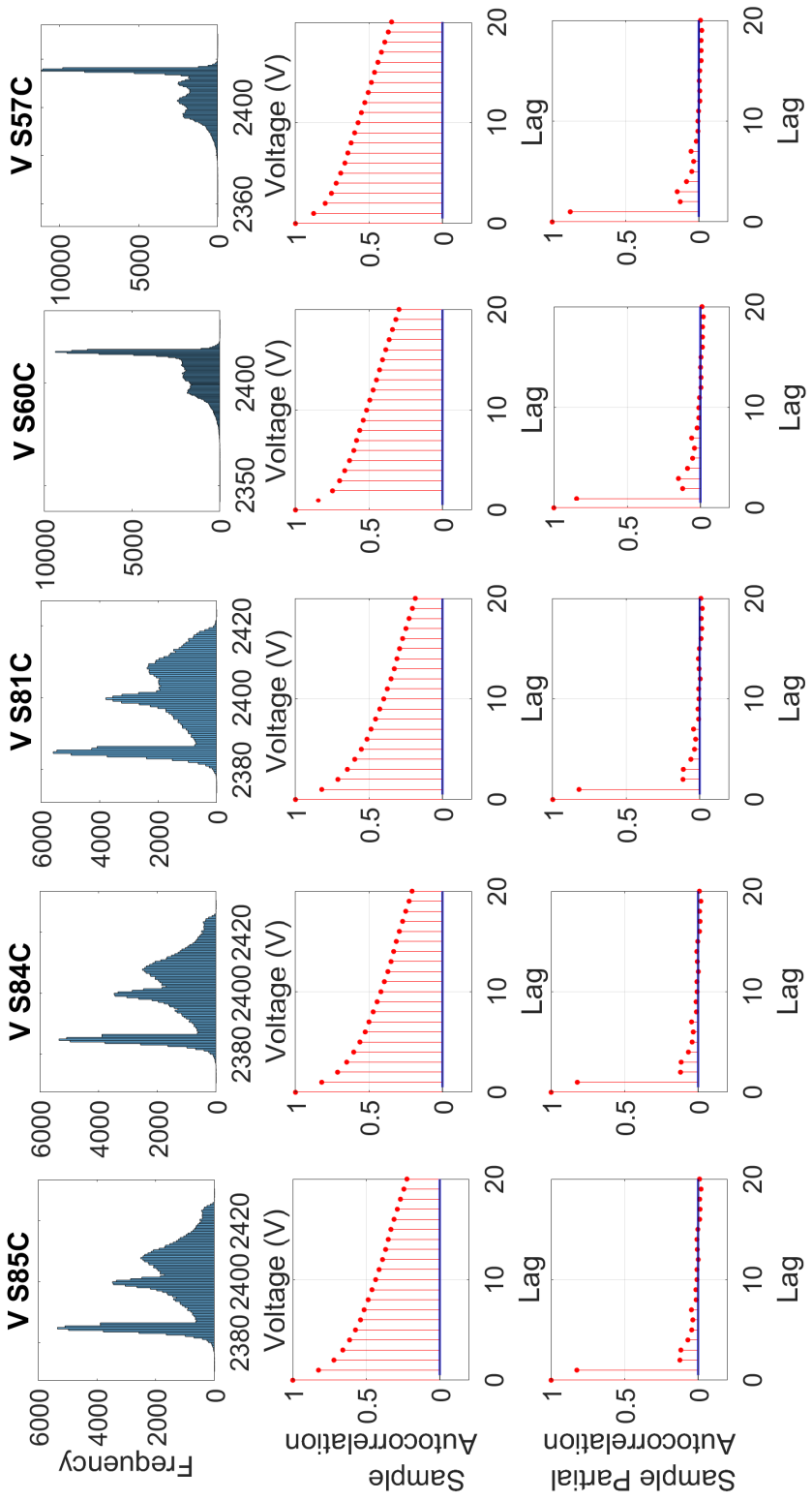


Figure 4.10: Distribution shape, ACF and PACF results of voltage distributions obtained from the reference case, showing non-Gaussian shapes and non-stationarity

Table 4.3: Tests results from measured voltages (on phase C)

Node	Test rejection	p-values	Test statistics DF_τ
S85.C	Failure to reject H_0	0.4545	-0.5392
S84.C	Failure to reject H_0	0.4635	-0.5143
S81.C	Failure to reject H_0	0.4761	-0.4801
S60.C	Failure to reject H_0	0.5016	-0.4104
S57.C	Failure to reject H_0	0.5276	-0.3395

As a result, a p-value helps infer the time series' stationarity. The testing procedure for the ADF test is applied to the model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{4.8}$$

where α is a constant, β is the coefficient on a time trend, and p is the lag order of the autoregressive process. A random walk is modelled by setting the constraints $\alpha = 0$ and $\beta = 0$, and making only $\beta = 0$ corresponds to modelling a random walk with a drift. Higher-order autoregressive processes are allowed when lags of the order p are included in the ADF formulation. Different ways of testing then include testing down from high orders lag length p and examine the t-values on coefficients or examining information criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC).

The unit root test is then conducted under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. The test statistic is computed and then compared with the relevant critical value for the Dickey-Fuller test, which follows a specifically known distribution as the Dickey–Fuller table for critical values. Calculation of test value is obtained from expression

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{4.9}$$

If the calculated test statistic is less than the critical value, then the null hypothesis $\gamma = 0$ is rejected and no unit root is present.

The results of applying the test over the data are shown in Table 4.3 (this values are only for phase C, however, this results applies for all phases), and it shows for each dataset the test rejection decisions for null hypothesis, the Test statistic p-values (which was the one sample t-test for left-tail probabilities) and the test statistic DF_τ.

4.4.1.2 First regression attempt and exploring of all data

From results obtained in Table 4.3, it is concluded that the data obtained requires to be differenced to make it stationary. Normally, applying data differencing of order one is enough to become stationary [246]. It can be an iterative process for several order until stationarity is achieved. However, a good revision of the data and the model would be suggested to give an interpretation to the obtained models and results.

The data presented before is now differenced and not absolute values. Once this data is stationary, it is desired to build an "good" LTI model from the statistical point of view. To achieve that, it is assumed (and desired) that the input data follows characteristics such as normality in the distribution's functions (a straightforward way to prove independence between variables).

A first preliminary regression is done in order to explore the residuals. The techniques that were explored in this thesis can be grouped as follow:

4.4.1.3 Linear autoregressive models

This approach has been used to obtain empirical models that relates inputs u_t and outputs y_t [102]. Most of these models are based on the equation error model structure, which is the simplest input-output relationship for stochastic processes, and the current output is product of the interaction between previous inputs and outputs. The family of possible models is big (about 32 possible model structure sets), but only ARXs and ARMAXs models will be discussed as the traditional tools used to provide a parsimonious description of the (weakly) stationary stochastic process. The most basic representation corresponds to the ARX model represented by

$$y_t + \sum_{i=1}^{n_a} a_i y_{t-i} = \sum_{j=1}^{n_b} b_j u_{t-j} + e_t \tag{4.10}$$

where e_t is white noise and the values of a_i and b_j are adjustable parameters that are found using linear regression. The name of this model comes from using autoregressive components associated with the outputs, plus the participation of exogenous variables associated with the inputs.

Previous model can be expanded considering the participation of the disturbance in the system summarised in e_t. Then, the ARMAX model is described as follow:

$$y_t + \sum_{i=1}^{n_a} a_i y_{t-i} = \sum_{j=1}^{n_b} b_j u_{t-j} + e_t + \sum_{k=1}^{n_c} c_k e_{t-k} \tag{4.11}$$

This addition increases flexibility in the model and describes the error as a moving

an average of white noise.

4.4.1.4 Koopman-operator-based models

Models presented before are based on a classical statistical and geometric perspectives on dynamical systems. There are other approaches based on the evolution of measurements of the system. The Koopman operator theory introduces a third operator-theoretic perspective, in which is expected to take advantage on the availability of measurement data from complex systems [136]. Koopman theory looks to the identification of intrinsic coordinate of a system to get a linear framework of its nonlinear dynamics [263]. Considering the original quasi-dynamic system presented in (3.1), the state of the system $x \in M$, where M is a differentiable manifold, often given by $M = \mathbb{R}^n$. The quasi-dynamics of the discrete equivalent system are given by $x_{k+1} = \mathbf{F}(x_k)$, where $\mathbf{F}$ may be the flow map of the quasi-dynamics of (3.1). A Koopman operator κ_t represents an infinite-dimensional linear operator that advances measurement functions of the state $g : M \to \mathbb{R}$ with the flow $\mathbf{F}$ of the quasi-dynamics given by

$$\kappa_t g = g \circ \mathbf{F} \tag{4.12}$$

One of the interesting and promising properties of this operator is its linearity in his infinite dimensional representation, which produce problem in its computation [31]. Therefore, it is expected to apply the Koopman analysis to approximates the evolution on a subspace spanned by a finite-dimensional set of measurement functions to an invariant subspace, instead of capturing the evolution of all measurement functions in a Hilbert space. As in any linear representation, a Koopman invariant subspace is spanned by any set of eigenfunctions $\gamma(x)$ of the Koopman operator, corresponding to eigenvalue λ, and it satisfies

$$\kappa_t \gamma = \lambda \gamma \tag{4.13}$$

Obtaining these eigenfunctions from data or analytically is challenging in general but discovering these eigenfunctions enables globally linear representations of strongly nonlinear systems in terms of these intrinsic observables [143]. One solution for this is a method called Dynamic Mode Decomposition or DMD, which is a simple numerical algorithm that approximates the Koopman operator [139]. This method consists of a modal decomposition, where each mode consists of spatially correlated structures that have the same linear behaviour in time. This produces the best-fit linear dynamical system that advances high-dimensional measurements forward in time [140]. The infinite-dimensional Koopman operator is

approximated with a finite-dimensional matrix A that advances the system state x:

$$x_{k+1} \approx Ax_k \tag{4.14}$$

Using data that represents the non-linear system, it is possible to represent the system using "snapshots" stored in matrices $X = [x_1, x_2 \cdots x_{m-1}]^T$ and time-shifted $X^{'} = [x_2, x_3 \cdots x_m]^T$. Equation (4.14) can be represented as follows:

$$X^{'} \approx AX \tag{4.15}$$

The DMD algorithm tries to find the leading eigendecomposition of the best-fit linear operator, given by

$$A = \arg\min_{A^*} \|X^{'} - A^*X\|_F \tag{4.16}$$

where $\|\cdot\|_F$ is the Frobenius norm. The best-fit A is given by $A = X^{'}X^{\dagger}$, where $\dagger$ is the pseudo-inverse, which is computed via Singular Value Decomposition (SVD).

This is how DMD approximates the Koopman operator restricted to the set of direct measurements of the state of a high-dimensional system. In this case, it was developed a Dynamic Mode Decomposition with Control (DMDc) to integrate the exogenous variables that are used to explain the voltage quasi-dynamic, and potentially used for control applications [132]. This extension proposed by Proctor et al. [264], consider the participation of natural unforced dynamics and the effect of actuation. In an equivalent way as shown in (4.15), the quasi-dynamics are represented by

$$x_{k+1} \approx Ax_k + Bu_k \tag{4.17}$$

Therefore, to obtain matrices A and B it is required also the data vectors X, $X^{'}$ and an actuation history matrix $\Upsilon = [u_1, u_2 \cdots u_m]^T$. Therefore, Equation (4.17) can be represented as follows:

$$X^{'} \approx AX + B\Upsilon \tag{4.18}$$

When matrix B is unknown (in this case, it is expected to obtain a reduced equivalent system quasi-dynamic), both matrices A and B must be calculated simultaneously. The approximation presented in (4.18) is reorganised as follows

$$X^{'} \approx \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X \\ \mathrm{Y} \end{bmatrix} = G\Omega \tag{4.19}$$

The matrix G is obtained using a least-squares regression, given by:

$$G \approx X^{'}\Omega^{\dagger} \tag{4.20}$$

The high-dimensional matrix $\Omega = [X^{*}\mathrm{Y}^{*}]^{*}$ is approximated using SVD, as follows:

$$\Omega = \tilde{U}\tilde{\Sigma}\tilde{V}^{*} \tag{4.21}$$

where $\tilde{U} = [\tilde{U}_1^{*}\tilde{U}_2^{*}]^{*}$provides a reduced bases for the input space. On the other hand, a reduced basis for the output space $\hat{U}$ defines the value of $X^{'}$ as follows:

$$X^{'} = \hat{U}\hat{\Sigma}\hat{V}^{*} \tag{4.22}$$

Then, the matrix G can be approximated by projecting onto this basis:

$$\tilde{G} = \hat{U}^{*}G \begin{bmatrix} \hat{U} \\ I \end{bmatrix} \tag{4.23}$$

The resulting projected matrices $\tilde{A}$ and $\tilde{B}$ in $\tilde{G}$ are calculated as follows:

$$\tilde{A} = \hat{U}^{*}A\hat{U} = \hat{U}^{*}X^{'}\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}_1^{*}\hat{U} \tag{4.24a}$$

$$\tilde{B} = \hat{U}^{*}B = \hat{U}^{*}X^{'}\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}_2^{*} \tag{4.24b}$$

Normally, the tuning is focused only on the dimension of first matrix $X^{'}$ presented in (4.22).

4.4.1.5 Subspace identification methods

Another technique used was the subspace identification method called Observed/Kalman Filter Identification and Eigensystem Realization Algorithm (OKID-ERA), which both algorithms constitute a classic system identification technique used in several applications. Nevertheless, the previous modal decomposition method have been shown to be intimately connected to OKID-ERA [265]. These two different algorithms are complemented to the system identification purpose: the first part produces a de-noised linear impulse response from the input, while the

second oversees producing a reduced-order state-space system[132].

The ERA algorithm uses impulse response data and produces low-dimension linear input–output models based on the “minimal realisation" theory, in which a Hankel matrix H is produced by stacking shifted time-series of impulse response measurements. The resulting is a low order defined by the numerical rank of the controllability and observability subspaces [102].

To explain how ERA algorithm works, consider the discrete-time system with a time-step k:

$$\overline{x}_{k+1} = A_d\overline{x}_k + B_d u_k \tag{4.25a}$$

$$y_k = C_d\overline{x}_k + D_d u_k \tag{4.25b}$$

A discrete-time delta function input in the actuation gives rise to a discrete-time impulse response in the sensors, as shown in the equations (4.26) and (4.27), respectively:

$$u_k^\delta \triangleq u^\delta(k\Delta t) = \begin{cases} I & k = 0 \\ 0 & k = 1, 2, 3 \cdots \end{cases} \tag{4.26}$$

$$y_k^\delta \triangleq y^\delta(k\Delta t) = \begin{cases} D_d & k = 0 \\ C_d A_d^{k-1} B_d & k = 1, 2, 3 \cdots \end{cases} \tag{4.27}$$

This interaction applying several inputs p is typically done one for each of the separate input channels q. The output responses are collected for each impulsive input at a given time-step k, which will produce a Hankel matrix. In fact, the presented matrices A_d,B_d,C_d and D_d could exist or not, since this method is data-driven and these matrices are representing what is represented by the Hankel matrix, which is formed by stacking shifted time-series impulse-response from measurements, as shown in the following expression:

$$
\begin{aligned}
H &= \begin{bmatrix} y_1^\delta & y_2^\delta & \cdots & y_{m_c}^\delta \\ y_2^\delta & y_3^\delta & \cdots & y_{m_c+1}^\delta \\ \vdots & \vdots & \ddots & \vdots \\ y_{m_o}^\delta & y_{m_o+1}^\delta & \cdots & y_{m_c+m_o-1}^\delta \end{bmatrix} \\
&= \begin{bmatrix} C_d B_d & C_d A_d B_d & \cdots & C_d A_d^{m_c-1} B_d \\ C_d A_d B_d & C_d A_d^2 B_d & \cdots & C_d A_d^{m_c} B_d \\ \vdots & \vdots & \ddots & \vdots \\ C_d A_d^{m_o-1} B_d & C_d A_d^{m_o} B_d & \cdots & C_d A_d^{m_c+m_o-2} B_d \end{bmatrix}
\end{aligned} \tag{4.28}
$$

Since the equivalent matrices of system are not required to be accessed, a shifted Hankel matrix $H^{'}$ can be used instead:

$$
\begin{aligned}
H^{'} &= \begin{bmatrix} y_2^\delta & y_3^\delta & \cdots & y_{m_c+1}^\delta \\ y_3^\delta & y_4^\delta & \cdots & y_{m_c+2}^\delta \\ \vdots & \vdots & \ddots & \vdots \\ y_{m_o+1}^\delta & y_{m_o+2}^\delta & \cdots & y_{m_c+m_o}^\delta \end{bmatrix} \\
&= \begin{bmatrix} C_d A_d B_d & C_d A_d^2 B_d & \cdots & C_d A_d^{m_c} B_d \\ C_d A_d^2 B_d & C_d A_d^3 B_d & \cdots & C_d A_d^{m_c+1} B_d \\ \vdots & \vdots & \ddots & \vdots \\ C_d A_d^{m_o} B_d & C_d A_d^{m_o+1} B_d & \cdots & C_d A_d^{m_c+m_o-1} B_d \end{bmatrix}
\end{aligned} \tag{4.29}
$$

Based on the matrices H and $H^{'}$, it is possible to construct a reduced-order model by detecting the dominant temporal patterns obtained from the SVD of H:

$$
H = U\Sigma V^* = \begin{bmatrix} \tilde{U} & U_t \end{bmatrix} \begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & \Sigma_t \end{bmatrix} \begin{bmatrix} \tilde{V}^* \\ V_t^* \end{bmatrix} \approx \tilde{U}\tilde{\Sigma}\tilde{V}^* \tag{4.30}
$$

Therefore, the reduced model system that can be obtained as follows:

$$
\tilde{x}_{k+1} = \tilde{A}\tilde{x}_k + \tilde{B}\tilde{u} \tag{4.31a}
$$

$$
y = \tilde{C}\tilde{x}_k \tag{4.31b}
$$

where

$$\tilde{A} = \tilde{\Sigma}^{-1/2}\tilde{U}^{*}H^{'}\tilde{V}\tilde{\Sigma}^{-1/2} \tag{4.32a}$$

$$\tilde{B} = \tilde{\Sigma}^{1/2}\tilde{V}^{*}\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} \tag{4.32b}$$

$$\tilde{C} = \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}\tilde{U}\tilde{\Sigma}^{1/2} \tag{4.32c}$$

It is not required some knowledge from the system; however, this algorithm is based on impulse response measurements which is not normally available in power system applications. The OKID algorithm works as a complement, since it approximates the impulse response from arbitrary input–output data. This algorithm uses an asymptotically stable Kalman filter to identify the Markov parameters of a system augmented [133, 266, 267]. These parameters are extracted from the observer Markov parameters, which approximate the impulse response of the system and can be used as inputs to the ERA algorithm.

4.4.1.6 Non-Linear regression (NN-based)

Finally, Nonlinear Auto-Regressive Model with Exogenous Inputs Mode (NARX) was used to contrast the previous results with a non-linear approach as reference of performance [144], which is a recurrent dynamic network with feedback connections enclosing several layers of a Neural Network (NN) used to map the non-linear components, and then referred to a regular ARX model structure. This method has been closely related with the Koopman-operator-based method SINDy, since both identifies the structure of models from time-series data through an orthogonal least square procedure [31, 268].

4.4.1.7 System State Identification

The previous linear models can be represented in state-space form, which is the main goal in this thesis. In general, linear structures can be represented in State-Space form [102, 130]. For example, most of linear models can be represented using the general-linear polynomial model or the general-linear model by the expression:

$$y(t) = G(q^{-1})u(t) + H(q^{-1})e(t) \tag{4.33}$$

where G and H are transfer functions in the time delay operator q^{-1}. This

general representation can be written as follows:

$$A(q^{-1})y(t) = q^{-d}\frac{B(q^{-1})}{F(q^{-1})}u(t) + \frac{C(q^{-1})}{D(q^{-1})}e_0(t) \tag{4.34}$$

where d is some multiple of the sampling period. Linear models such as ARX or ARMAX are particular model structures of this representation. To illustrate the equivalence between this representation with State-Sapce representation, ARX will be used as an example, in which the transfer function G and H are defined using parameters θ by the expressions:

$$G(q^{-1},\theta) = q^{-d}\frac{B_{ARX}(q^{-1})}{F_{ARX}(q^{-1})} \tag{4.35a}$$

$$H(q^{-1},\theta) = q^{-d}\frac{1}{A_{ARX}(q^{-1})} \tag{4.35b}$$

A simple relationship between the state space innovations and the general input-output form exists and it is given by the expressions:

$$G(q^{-1},\theta) = C_{SSIF}(\theta)[qI - A_{SSIF}(\theta)]^{-1}B_{SSIF}(\theta) \tag{4.36a}$$

$$H(q^{-1},\theta) = C_{SSIF}(\theta)[qI - A_{SSIF}(\theta)]^{-1}K_{SSIF}(\theta) + I \tag{4.36b}$$

Therefore, the deterministic part G of both expression can be rewritten as follows:

$$q^{-d}\frac{B_{ARX}(q^{-1})}{F_{ARX}(q^{-1})} = C_{SSIF}(\theta)[qI - A_{SSIF}(\theta)]^{-1}B_{SSIF}(\theta) \tag{4.37}$$

After some manipulations, it can be verified that the poles of the predictor are the eigenvalues from $A - KC$. The set D_m is given by the expression:

$$D_m = \{\theta | eig[A(\theta) - K(\theta)C(\theta)]\text{inside the unit circle}\} \tag{4.38}$$

When estimating state space models, the elements in $K(\theta)$ are typically estimated directly rather than using the detour of estimating the covariance matrices and solving the Ricatti equation before computing $K(\theta)$. Then the parameters A, B, C and K must be identified, which is not identifiable from input-output data. Therefore, the same input-output relationship can be described by different choices of A, B, C and K. Consequently, converting from input-output model to

state-space model is not done in a unique way, and there is not guarantee that the new state-space model will take all critical states to be observed/controlled. Nevertheless, there are advantages of this representation (e.g., easy implementation, take into account initial conditions, gives some insight of other properties of the system such as controllability and observability, this is a time domain method suitable for digital computer computation, among others), even if there is not unique way to make this conversion (in the case of linear autoregressive model) [104]. If it is not stated something different, all obtained models will be presented in their state-space form, as presented in equation (4.5).

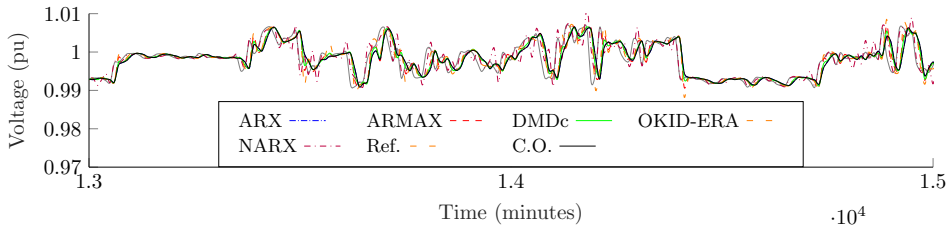
4.4.1.8 Obtained models and output results after first regressions

In this part of the approach, a MIMO structure is assumed to integrate the inputs and outputs. As this is an exploratory phase aimed at understanding the nature of the data and avoiding computational issues, the modelling is split into two parts. The first part comprises a model that includes variables from phases A and B, while the second part focuses on phase C. The obtained parameters for both parts of each model are presented in Appendix J.

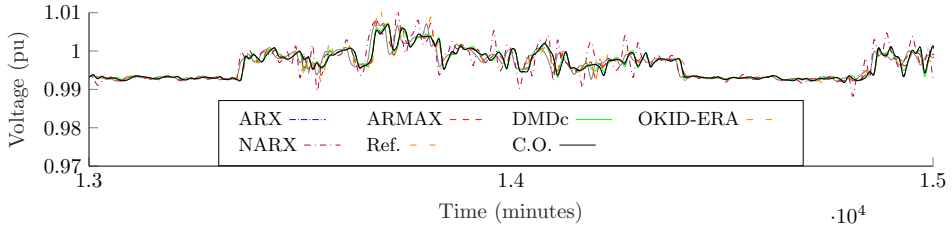
A portion of the results obtained from the first guess and the predicted voltages are presented in Figures 4.11 and 4.13, for all phases from the 50% of data used for training, while in Figures 4.12 and 4.14, for all phases from the 50% of data used for validation. These figures include the reference signals and the response from a simple persistent model (with constant output). Tables 4.4 and 4.5 summarise general characteristics of each obtained models. To make all models comparable regarding the capacity of representing the system quasi-dynamics using available data, all models were developed under similar conditions of dimension size (only the model obtained using DMDc is naturally reduced due to the algorithm simplification based on the SVD). It is not part of this thesis aim to get the optimal model for each algorithm. Therefore, it was not required a wide search space for models dimension.

Table 4.4: Obtained models dimensions for voltage prediction using raw training dataset (phase A and B)

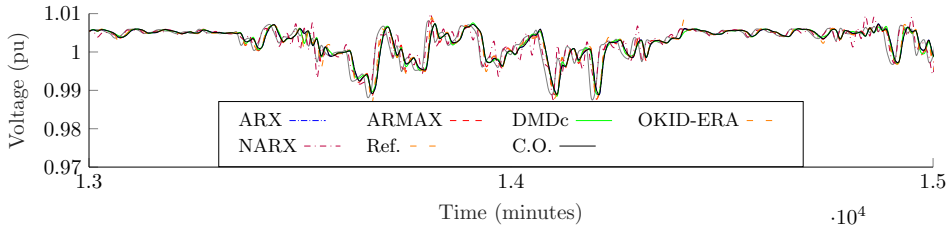
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension A	48x48	48x48	6x6	48x48	-
Dimension B	48x12	48x12	6x12	48x12	-
Dimension C	6x48	6x48	6x6	6x48	-
Dimension D	6x12	6x12	6x12	6x12	-
Comp. time (s)	122.37	30.98	0.32	1023.6	195.13



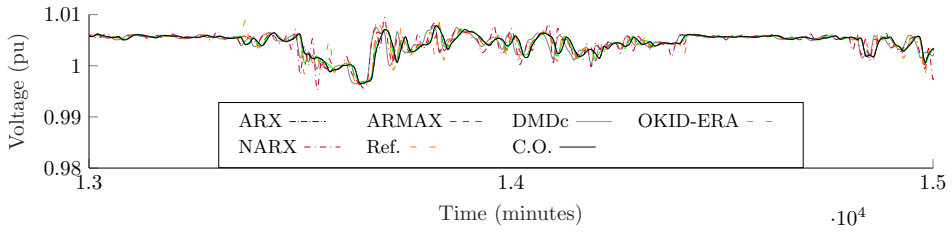
(a) Node S81A



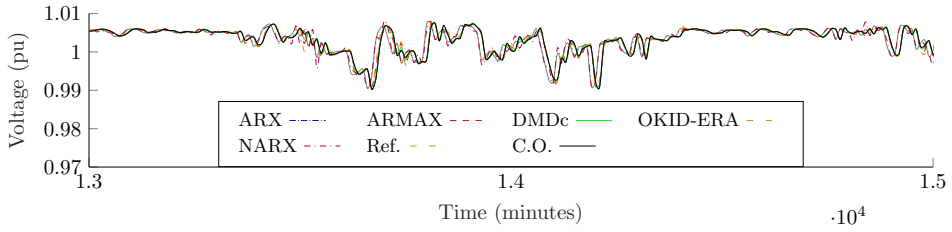
(b) Node S60A



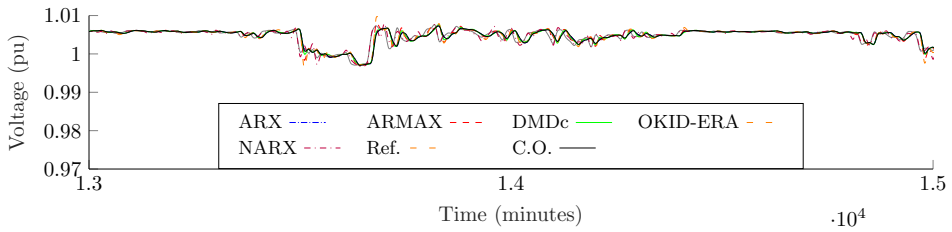
(c) Node S57A



(d) Node S81B

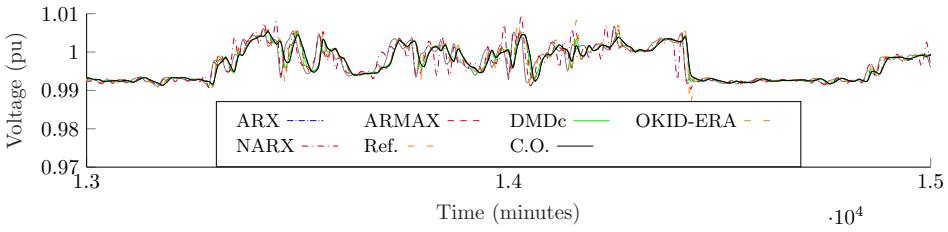


(e) Node S60B

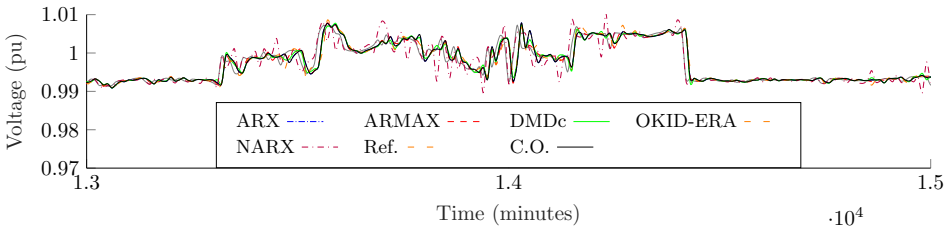


(f) Node S57B

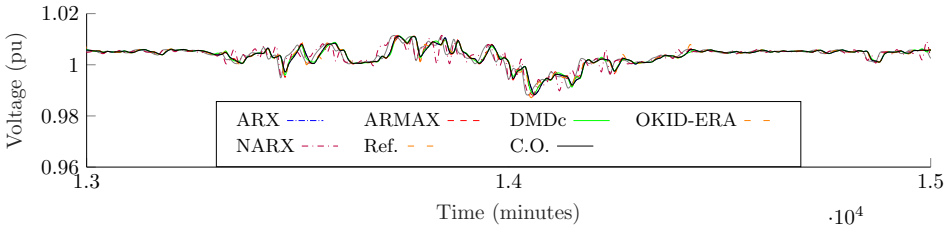
Figure 4.11: Portion of voltage predictions 1 step ahead using raw training dataset Phase A and B



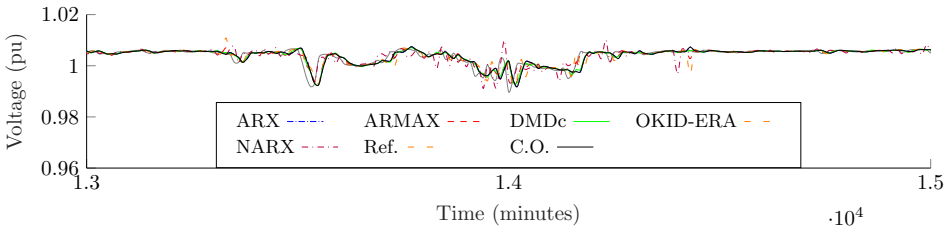
(a) Node S81A



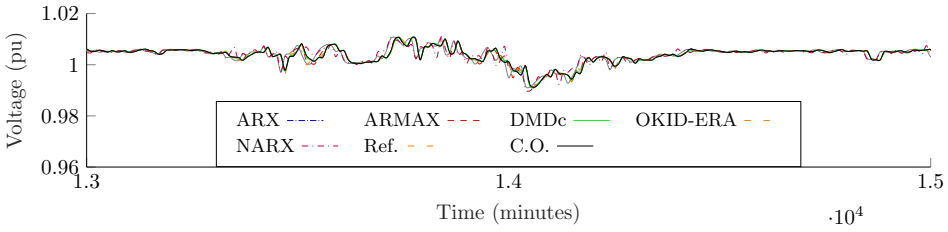
(b) Node S60A



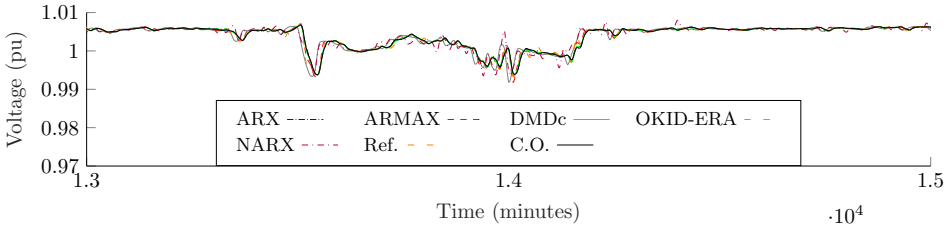
(c) Node S57A



(d) Node S81B

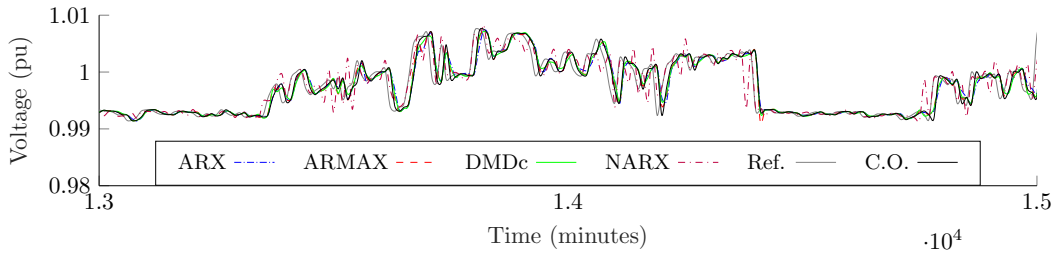


(e) Node S60B

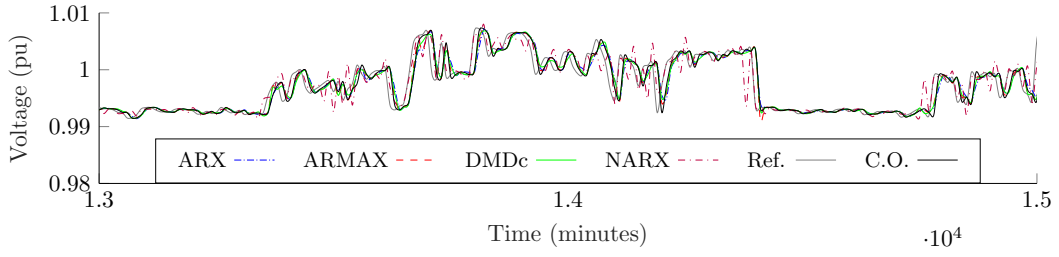


(f) Node S57B

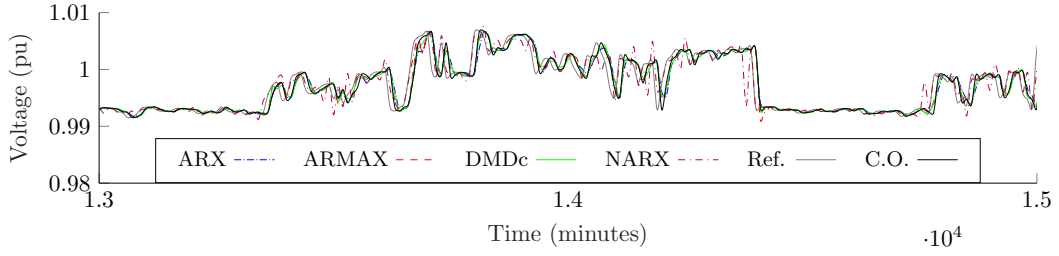
Figure 4.12: Portion of voltage predictions 1 step ahead using raw validation dataset Phase A and B



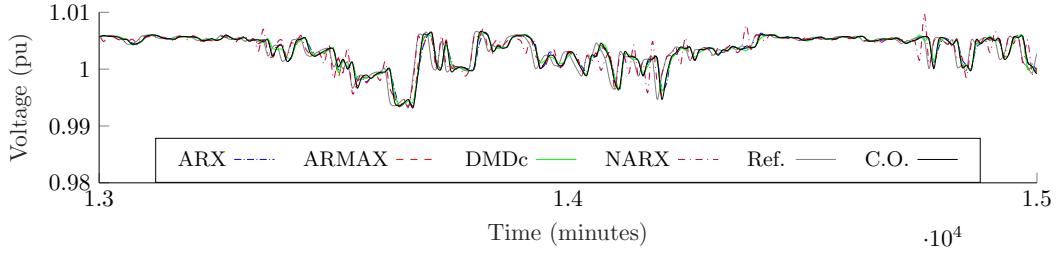
(a) Node S85C



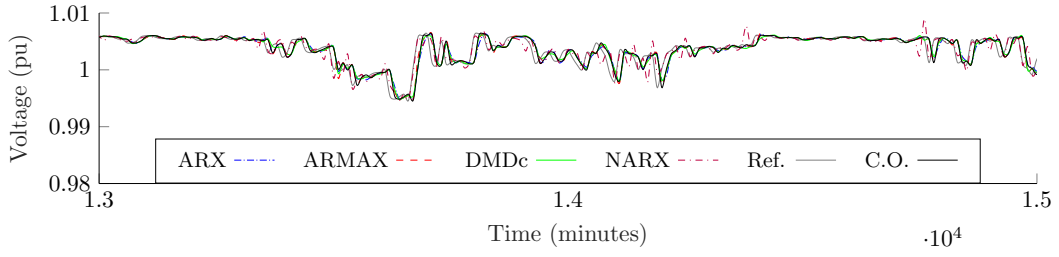
(b) Node S84C



(c) Node S81C

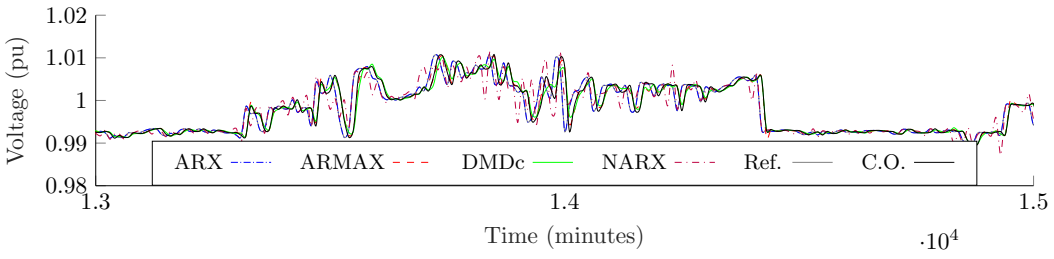


(d) Node S60C

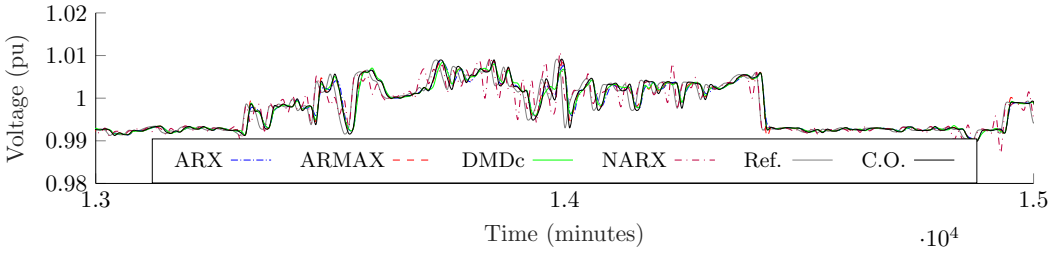


(e) Node S57C

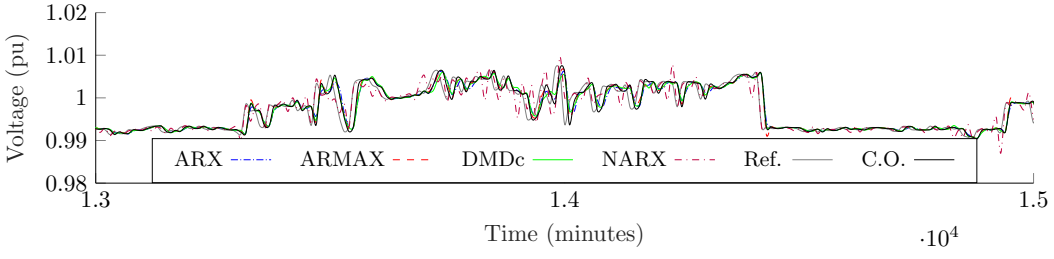
Figure 4.13: Portion of voltage predictions 1 step ahead using raw training dataset Phase C



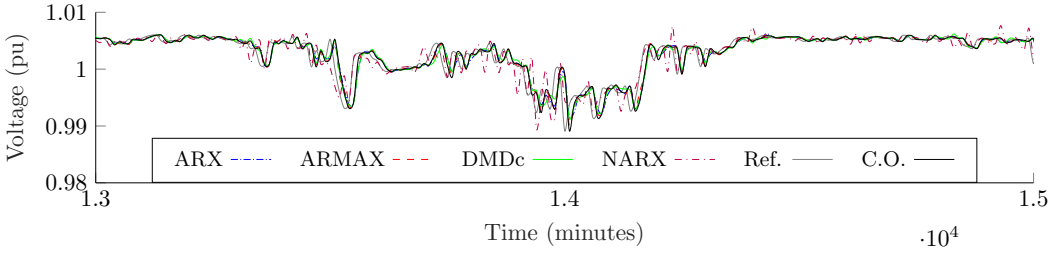
(a) Node S85C



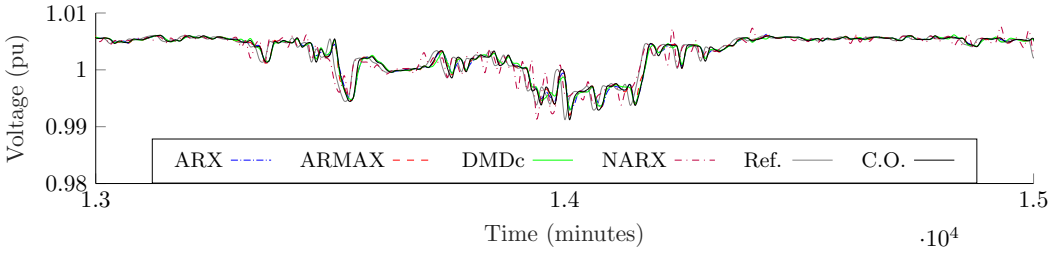
(b) Node S84C



(c) Node S81C



(d) Node S60C



(e) Node S57C

Figure 4.14: Portion of voltage predictions 1 step ahead using raw validation dataset Phase C

Table 4.5: Obtained models dimensions for voltage prediction using raw training dataset (phase C)

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension A	40x40	40x40	5x5	40x40	-
Dimension B	40x15	40x15	5x15	40x15	-
Dimension C	5x40	5x40	5x5	5x40	-
Dimension D	5x15	5x15	5x15	5x15	-
Comp. time (s)	124.94	46.84	0.77	740.03	252.46

ARX and ARMAX models are presented in their state-space (SS) representation. The autoregressive structures ARX, ARMAX and NARX consider output delays up to lag 7, and internal input delay up to lag 3. The first matrix $X^{'}$ for DMDc consider the matrix order equivalent of the ARX state-space representation in the exploration process, to have a broader exploration of relevant eigenvalues obtained after the SVD process. Then, the process reduces the matrix by up to the number of outputs. The OKID-ERA approach showed the best performance when the number of Markov parameters was set up to 20.

Tables 4.6 and 4.8 shows the performance for training and validation for the model of phases A and B, while Tables 4.7 and 4.9 shows the performance for training and validation for the model of phase C, respectively. Also, the response from the simple persistent model or constant output is used as reference in the obtained metrics. These metrics to compare models performance are commonly used in similar works. The R-squared (R^2) is a statistical measure to compare the data to the fitted regression line, which is the percentage of the response variable variation that is explained by a linear model, as shown in the following equation:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{4.39}$$

where $SS_{\text{res}} = \sum_i^n (y_i - f_i)^2$ is the sum of squares of residuals, y_i is the observed data and f_i is the fitted data and $SS_{\text{tot}} = \sum_{i=1}^n (y_i - \bar{y_m})^2$ is the total sum of squares. This metric gives an estimate of the relationship between movements of a dependent variable based on an independent variable's movements, but it doesn't say anything whether the data and predictions are biased, nor whether how good the model is. That is why it is required to use additional metrics to have a better understanding of the obtained models.

The root-mean-square error (RMSE) is a used to represent the square root of the quadratic mean of the differences between predicted and observed values. This metric is commonly normalised to facilitate the comparison between models with

Table 4.6: Results of models for voltage prediction using raw training dataset, phases A and B

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX	Constant O.
R^2 S81A	0.51	0.51	0.53	0.44	0.75	0.44
R^2 S60A	0.68	0.67	0.69	0.65	0.74	0.60
R^2 S57A	0.62	0.61	0.63	0.59	0.83	0.54
R^2 S81B	0.69	0.71	0.72	0.58	0.76	0.62
R^2 S60B	0.64	0.64	0.65	0.63	0.92	0.54
R^2 S57B	0.81	0.80	0.81	0.78	0.93	0.74
NRMSE S81A	0.14	0.14	0.13	0.15	0.10	0.21
NRMSE S60A	0.13	0.13	0.12	0.13	0.11	0.21
NRMSE S57A	0.08	0.08	0.08	0.08	0.05	0.16
NRMSE S81B	0.06	0.05	0.05	0.06	0.05	0.13
NRMSE S60B	0.09	0.09	0.09	0.09	0.04	0.18
NRMSE S57B	0.05	0.05	0.05	0.06	0.03	0.12
AIC	-3.825e5	-3.821e5	-3.835e5	-3.784e5	-4.051e5	-3.83e5
BIC	-3.824e5	-3.821e5	-3.835e5	-3.783e5	-4.051e5	-3.82e5

different scales, called normalised root-mean-square error (NRMSE), calculated using the following expression:

$$\text{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(f_{ij}-y_i)^2}{n}}}{y_{\max} - y_{\min}} \tag{4.40}$$

NRMSE perform well if the response variable is log-transformed, standardised, or otherwise modified, or if comparing models fits for different response variables. However, the downside is that it is lost the units associated with the response variable. Additional metrics are used to compare model relative performances: AIC and BIC, which are given according to the following equations, respectively:

$$\text{AIC} = n\log(SS_{\text{res}}/n) + 2p, \tag{4.41}$$

$$\text{BIC} = n\log(SS_{\text{res}}/n) + 2(p+2)s - 2q2, \tag{4.42}$$

where p is the number of model parameters, $q = n\sigma^2/SS_{\text{res}}$, and σ^2 is an estimate of the pure error variance from fitting the full model. The selection of correct model is a complex balance of checking all these metrics, which are used in this thesis to compare relative performances between models for the prepared data.

So far, the models are done using the inputs and outputs presented in Section 4.3. These results show that most of the methods for system identification used were suitable for the problem and the variables used in modelling. All methods except for OKID-ERA were able to produce a better performance in comparison

Table 4.7: Results of models for voltage prediction using raw training dataset, phase C

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX	Constant O.
R^2 S85C	0.67	0.67	0.68	-254.13	0.77	0.60
R^2 S84C	0.66	0.66	0.67	-284.03	0.79	0.60
R^2 S81C	0.66	0.66	0.67	-325.23	0.80	0.60
R^2 S60C	0.70	0.70	0.71	-302.73	0.83	0.61
R^2 S57C	0.76	0.76	0.77	-246.06	0.87	0.69
NRMSE S85C	0.11	0.11	0.11	2.98	0.09	0.18
NRMSE S84C	0.11	0.11	0.11	3.27	0.09	0.18
NRMSE S81C	0.12	0.12	0.12	3.82	0.10	0.18
NRMSE S60C	0.06	0.06	0.06	1.82	0.04	0.12
NRMSE S57C	0.06	0.06	0.05	1.80	0.04	0.12
AIC	-3.789e5	-3.789e5	-3.8e5	-1.654e5	-3.946e5	-3.78e5
BIC	-3.789e5	-3.789e5	-3.8e5	-1.654e5	-3.945e5	-3.78e5

Table 4.8: Results of models for voltage prediction using raw validation dataset, phases A and B

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX	Constant O.
R^2 S81A	0.51	0.51	0.52	0.43	0.74	0.44
R^2 S60A	0.66	0.65	0.67	0.63	0.73	0.59
R^2 S57A	0.61	0.59	0.62	0.57	0.82	0.54
R^2 S81B	0.68	0.69	0.70	0.56	0.74	0.61
R^2 S60B	0.62	0.62	0.63	0.61	0.91	0.55
R^2 S57B	0.80	0.79	0.80	0.77	0.92	0.73
NRMSE S81A	0.13	0.13	0.13	0.14	0.10	0.21
NRMSE S60A	0.13	0.13	0.13	0.14	0.12	0.21
NRMSE S57A	0.09	0.09	0.09	0.10	0.06	0.17
NRMSE S81B	0.06	0.05	0.05	0.06	0.05	0.14
NRMSE S60B	0.10	0.10	0.10	0.10	0.05	0.17
NRMSE S57B	0.05	0.05	0.05	0.06	0.03	0.12
AIC	-3.816e5	-3.811e5	-3.826e5	-3.773e5	-4.038e5	-3.81e5
BIC	-3.815e5	-3.811e5	-3.825e5	-3.773e5	-4.038e5	-3.82e5

the the simple persistent model. The worst performance was obtained from the OKID-ERA algorithm, which is quite sensitive during the tuning of parameters (in this case, the balance between the observer Markov parameter dimension and the identified system order). For this reason, the results from OKID-ERA were not plotted from Figures 4.13 and 4.14. Among the linear approaches tested, the best performance was obtained from the DMDc method. It outperformed other linear approaches such as ARX and ARMAX in terms of performance indicators and computation time. The DMDc method provided more favourable results, with smaller dimensions of the obtained matrices and significantly shorter computation time. On the other hand, it was expected that the NARX method would show better performance due to its ability to capture nonlinear dynamics. However, the

Table 4.9: Results of models for voltage prediction using raw validation dataset, phase C

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX	Constant O.
R^2 S85C	0.66	0.66	0.67	-265.98	0.77	0.59
R^2 S84C	0.66	0.66	0.67	-298.46	0.78	0.59
R^2 S81C	0.65	0.65	0.67	-343.79	0.79	0.59
R^2 S60C	0.69	0.69	0.70	-320.59	0.83	0.62
R^2 S57C	0.76	0.75	0.76	-261.24	0.87	0.69
NRMSE S85C	0.11	0.11	0.10	2.99	0.09	0.18
NRMSE S84C	0.11	0.11	0.11	3.29	0.09	0.18
NRMSE S81C	0.12	0.12	0.12	3.90	0.10	0.20
NRMSE S60C	0.08	0.08	0.07	2.47	0.06	0.15
NRMSE S57C	0.07	0.07	0.07	2.34	0.05	0.14
AIC	-3.784e5	-3.783e5	-3.7950e5	-1.641e5	-3.946e5	-3.77e5
BIC	-3.784e5	-3.782e5	-3.7945e5	-1.640e5	-3.945e5	-3.77e5

DMDc method proved to be more effective in this case. One disadvantage of the NARX method is its inability to incorporate features that describe the system's internal dynamics. This is a desired characteristic in traditional linear control approaches. Additionally, the NARX method is computationally expensive, which can make it challenging to integrate into a real-time control approach.

4.4.1.9 Evaluating the residuals after first regressions

For a linear model in standard conditions, a good regression produces residuals that follow a normal distribution function (white noise), and additionally, no autocorrelation nor heteroscedasticity components [246, 260]. Therefore, it is required to see the characteristics of the residuals obtained after the first approximation. As illustration, Figures 4.15 and 4.16 presents the histogram, the quantile-quantile plot or the Q-Q plot (a probability plot used in this case to compare the normal distribution function with the obtained probability distributions by plotting their quantiles against each other), and ACF components of the residuals for training and validation of DMDc in phase C, which showed the best performance in overall from the linear modelling approaches. Comparable results were obtained for the other methods and phases. Figures 4.17 and 4.18 presents the results for the residuals of NARX to compare the performance. It is shown for both methods graphically that the distribution residuals are heavy-tailed, non-Gaussian with autocorrelation components, which means that the error obtained is still depending on previous values. These were not part of the initial assumptions. These characteristics are explored using different techniques to make a more rigorous analysis.

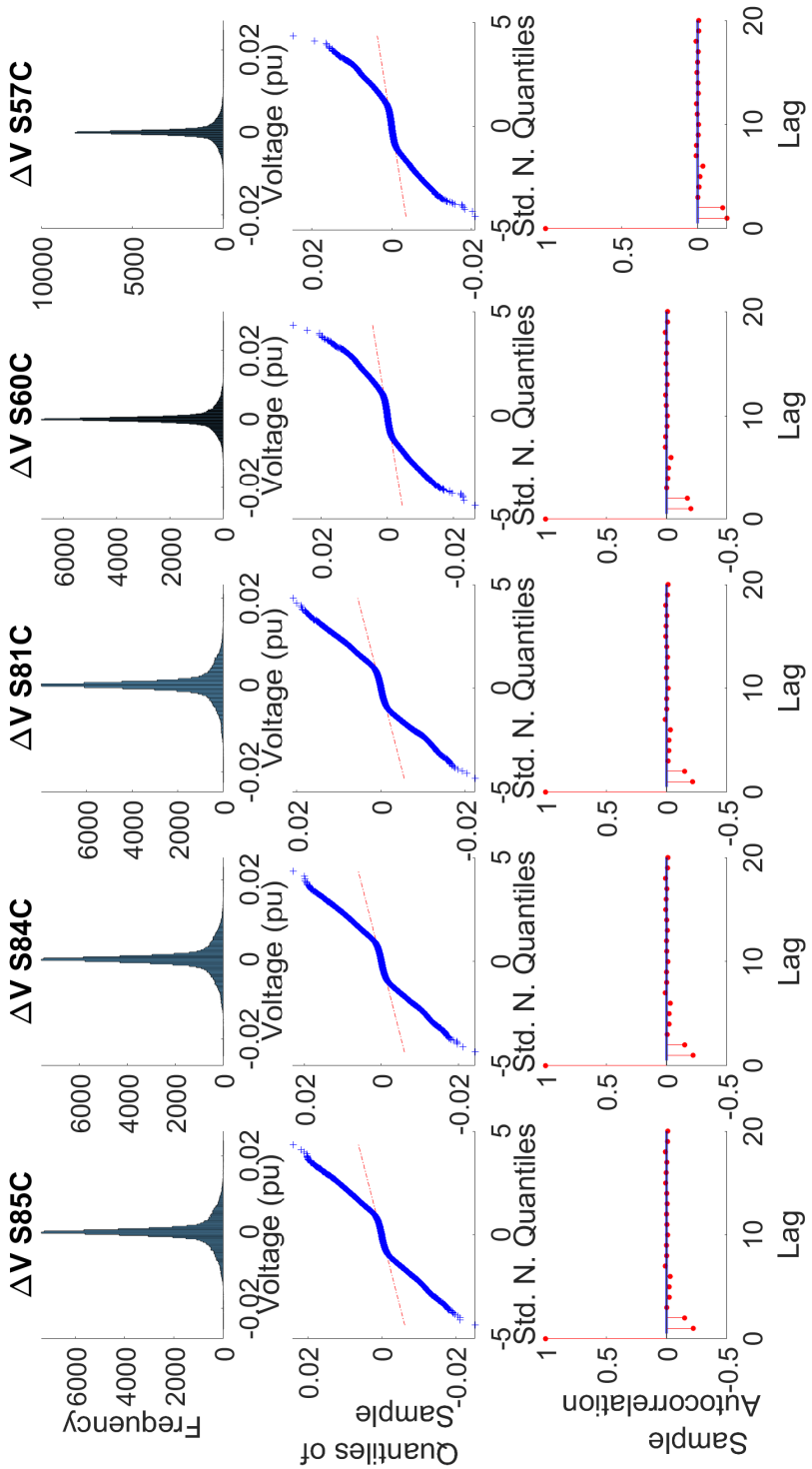


Figure 4.15: Histogram, Q-Q plot and ACF of residuals from voltages predictions on phase C using DMDc technique and raw training dataset

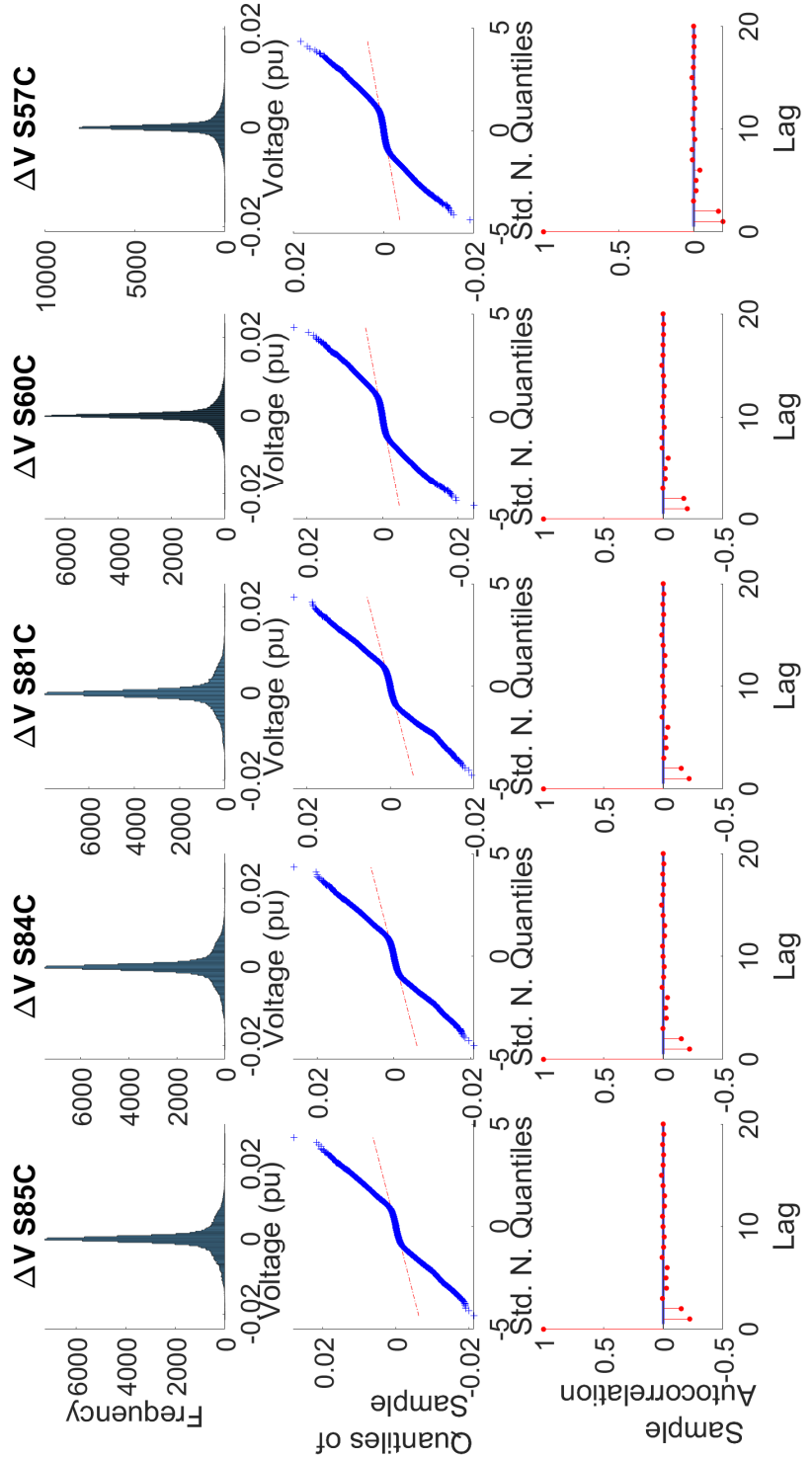


Figure 4.16: Histogram, Q-Q plot and ACF of residuals from voltages predictions on phase C using DMDc technique and raw validation dataset

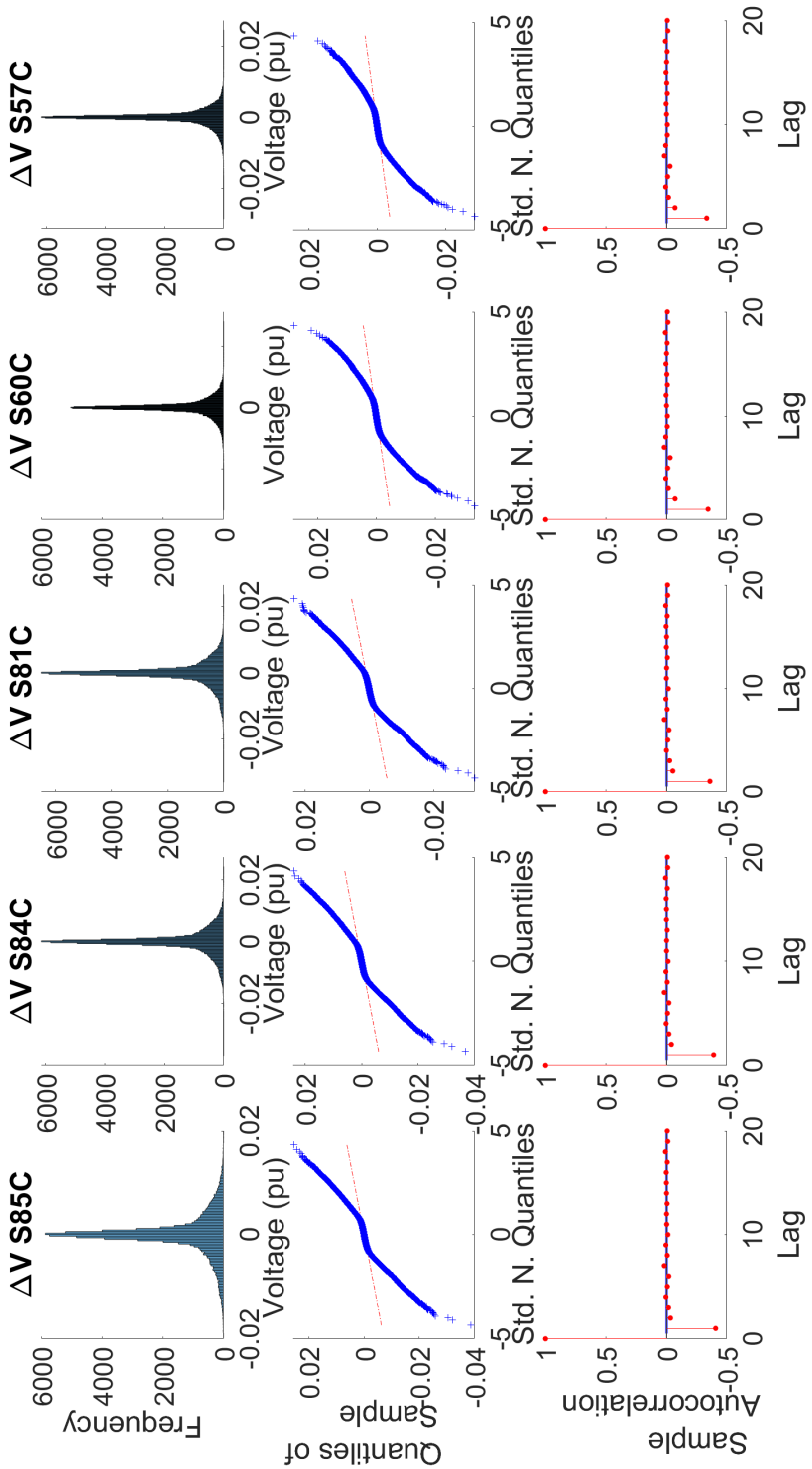


Figure 4.17: Histogram, Q-Q plot and ACF of residuals from voltages predictions on phase C using NARX technique and raw training dataset

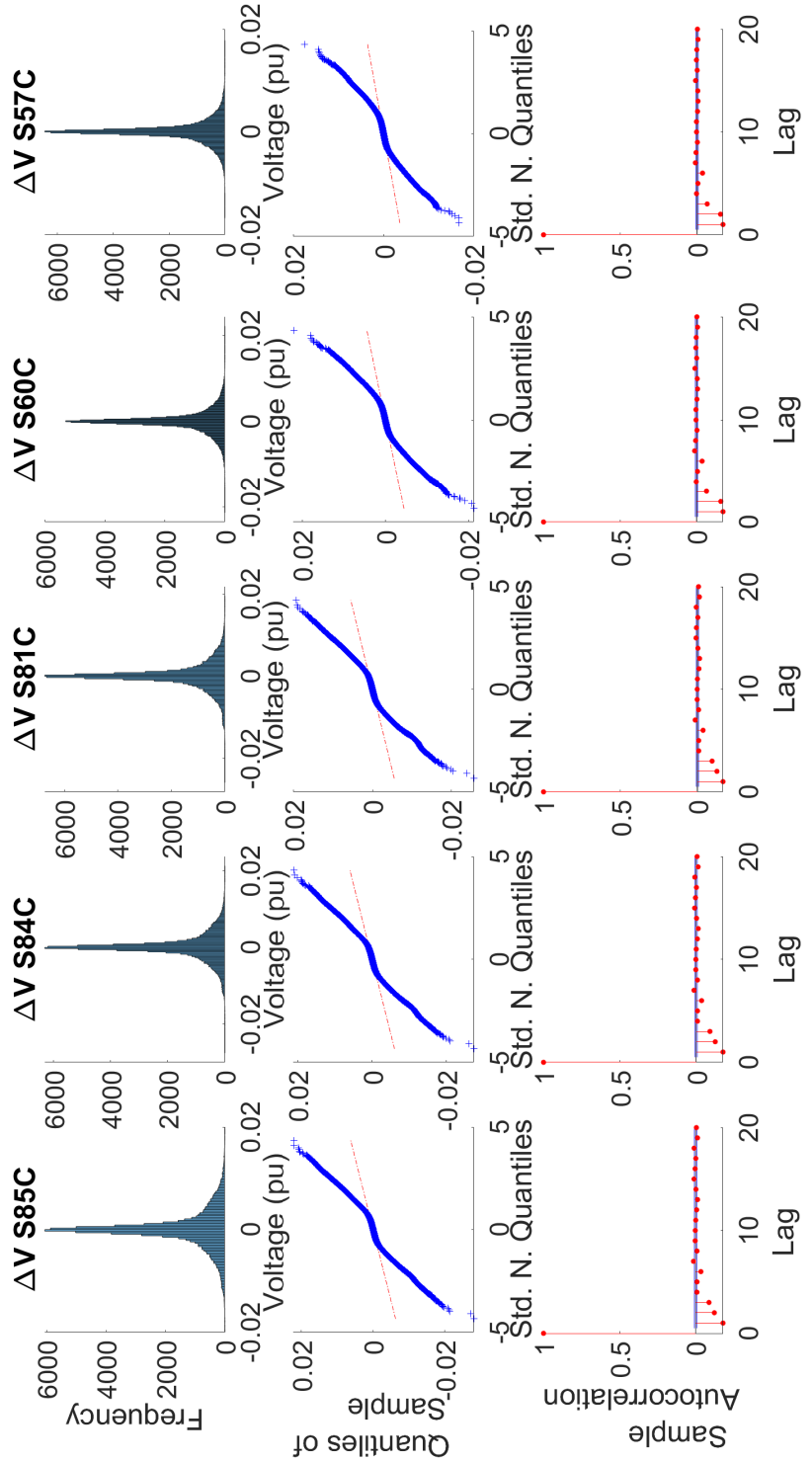


Figure 4.18: Histogram, Q-Q plot and ACF of residuals from voltages predictions on phase C using NARX technique and raw validation dataset

For the normality tests presented in this section, it is defined the following hypothesises:

Hypothesis H_0: The data follow a normal distribution.

Hypothesis H_1: The data do not follow the normal distribution.

- Anderson-Darling test [269, 270]:

  The Anderson–Darling test is based on empirical distribution function, more specifically to the quadratic empirical distribution function statistics, which measure the distance between the hypothesised distribution F (in this case, the Gaussian distribution) and empirical cumulative distribution function F_n, given by

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) \, dF(x), \tag{4.43}$$

  where n represents the sample size, and $w(x)$ is a weighting function. For this test, the weighting function is assumed as $w(x) = [F(x)\,(1 - F(x))]^{-1}$, and the test is based on the distance A^2, which is calculated as follows:

$$A^2 = n \int_{-\infty}^{\infty} \left(\frac{F_n(x) - F(x))^2}{F(x)\,(1 - F(x))}\right) dF(x), \tag{4.44}$$

  This Anderson–Darling distance gives more weight to all observations in the tails of the distribution, in comparison with other tests such as the Cramér–von Mises test.

  The Anderson–Darling test makes use of the fact that the CDF of the data can be assumed to follow a uniform distribution, when given the hypothesised underlying distribution and assuming the data does arise from this normal distribution. The test statistic is compared against its p value from the theoretical distribution with a significance level α. Additionally, it is not required to estimate any parameters in relation to the cumulative normal distribution function F.

  The formula for the test statistic A to assess if sorted data $\{x_1 < \cdots < x_n\}$ comes from a CDF of F is defined as follows:

$$A^2 = -n - S, \tag{4.45}$$

  where

$$S = \sum_{i=1}^{n} \frac{2i - 1}{n} \left[\ln(F(x_i)) + \ln\left(1 - F(x_{n+1-i})\right)\right] \tag{4.46}$$

- Cramer-Von Mises Test [271]:

  The Cramer-Von Mises Test works similar to the Anderson-Darling Test, in which is measured the distance between the hypothesised distribution F (in this case, the Gaussian distribution) and empirical cumulative distribution function F_n, given by

  $$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x), \tag{4.47}$$

  In this test, one of the functions F is the theoretical distribution and F_n is the empirically observed distribution, which is called one-sample case, while the two two-sample case occurs when both distributions are empirically estimated.

  The formula for the test statistic T to assess if sorted data $\{x_1 < \cdots < x_n\}$ comes from a CDF of F is defined as follows:

  $$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2 \tag{4.48}$$

  The test statistic is compared against its p value from the theoretical distribution with a significance level α.

- Shapiro-Wilk Test [270, 272]:

  The Shapiro-Wilk Test considers the randomly sorted data $\{x_1 < \cdots < x_n\}$ and evaluates the null hypothesis that this sample is represented by a normal distribution function. The test statistic W is calculated as follows:

  $$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{4.49}$$

  where $x_{(i)}$ is the ith order statistic, i.e., the ith-smallest number in the sample or the sorted vector of x (it is not the same as x_i), and $\overline{x}$ is the sample mean. The values of $(a_1, \ldots, a_n)$ are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution, and they are given by the formula:

  $$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{\|V^{-1} m\|} = \frac{m^{\mathsf{T}} V^{-1}}{(m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}} \tag{4.50}$$

  where V is the covariance matrix of those normal order statistics, and the vector $m = (m_1, \ldots, m_n)^{\mathsf{T}}$ is constructed using the expected values of the

order statistics of independent and identically distributed random variables sampled from the standard normal distribution. The distribution W has no name and the cut-off values for the statistics are calculated using Monte Carlo simulations. The test statistic is compared against its p value with a significance level α.

- D'Agostino and Pearson Test [273, 274]: This test is based on calculating the sample skewness and kurtosis and compare it with the reference of same values in a normal distribution function. The values are calculated as follows, respectively:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}\,, \tag{4.51}$$

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{2}} - 3 \tag{4.52}$$

where m_j are the j-th sample central moments, and $\bar{x}$ is the sample mean with a size n. The sample skewness g_1 and kurtosis g_2 are both asymptotically normal, but the rate of their convergence to the distribution limit is considerably slow. To improve this response, it is suggested to transform the values in a way that makes their distribution as close to standard normal distribution as possible. If the sample comes from a normal population, the exact finite sample distributions of the skewness and kurtosis can themselves be analysed in terms of their means, variances, skewnesses, and kurtoses, which derives to the following expressions, respectively:

$$\mu_1(g_1) = 0, \tag{4.53a}$$

$$\mu_2(g_1) = \frac{6(n-2)}{(n+1)(n+3)}, \tag{4.53b}$$

$$\gamma_1(g_1) \equiv \frac{\mu_3(g_1)}{\mu_2(g_1)^{3/2}} = 0, \tag{4.53c}$$

$$\gamma_2(g_1) \equiv \frac{\mu_4(g_1)}{\mu_2(g_1)^2} - 3 = \frac{36(n-7)(n^2+2n-5)}{(n-2)(n+5)(n+7)(n+9)}. \tag{4.53d}$$

$$\mu_1(g_2) = -\frac{6}{n+1}, \tag{4.54a}$$

$$\mu_2(g_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}, \tag{4.54b}$$

$$\gamma_1(g_2) \equiv \frac{\mu_3(g_2)}{\mu_2(g_2)^{3/2}} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}, \tag{4.54c}$$

$$\gamma_2(g_2) \equiv \frac{\mu_4(g_2)}{\mu_2(g_2)^2} - 3 = \frac{36(15n^6-36n^5-628n^4+982n^3+5777n^2-6402n+900)}{n(n-3)(n-2)(n+7)(n+9)(n+11)(n+13)}. \tag{4.54d}$$

The transformation proposed for this test corresponds to the following equations:

$$Z_1(g_1) = \delta \operatorname{asinh}\left(\frac{g_1}{\alpha'\sqrt{\mu_2(g_1)}}\right), \tag{4.55a}$$

$$\delta = \frac{1}{\sqrt{\ln W'}}, \tag{4.55b}$$

$$\alpha'^2 = \frac{2}{(W'^2-1)}, \tag{4.55c}$$

$$W'^2 = \sqrt{2\gamma_2(g_1)+4} - 1. \tag{4.55d}$$

$$Z_2(g_2) = \sqrt{\frac{9A}{2}}\left\{1 - \frac{2}{9A} - \left(\frac{1-\frac{2}{A}}{1+\frac{g_2-\mu_1(g_2)}{\sqrt{\mu_2(g_2)}}\sqrt{\frac{2}{(A-4)}}}\right)^{1/3}\right\}, \tag{4.56a}$$

$$A = 6 + \frac{8}{\gamma_1(g_2)}\left(\frac{2}{\gamma_1(g_2)} + \sqrt{1+\frac{4}{\gamma_1(g_2)^2}}\right). \tag{4.56b}$$

(4.56c)

Finally, statistics $Z_1(g_1)$ and $Z_2(g_2)$ are combined to produce an omnibus test K^2, which is capable to detect deviations from normality due to either skewness or kurtosis, calculated as follows:

$$K^2 = Z_1(g_1)^2 + Z_2(g_2)^2 \tag{4.57}$$

The test statistic is compared against its p value with a significance level α.

- Kolmogorov–Smirnov (KS) Lilliefors Modification test [275, 276]:

  This test (based on Kolmogorov–Smirnov test) measures the maximum distance between the hypothesised distribution F (in this case, a normal distribution function) and the empirical cumulative distribution function F_n for the sorted data $\{x_1 < \cdots < x_n\}$, and it is given by

$$D^* = \max_x |F_n(x) - F(x)| \tag{4.58}$$

  This calculation is done by first estimating the population mean and variance of the data. the value D^* is calculated, which is the test statistic that is compared against its p value with a significance level α. The purpose of this method is to assess whether the maximum discrepancy is large enough to be statistically significant. The obtained distribution is called Lilliefors distribution, and tables for this distribution is computed only by Monte Carlo methods.

Autocorrelation in the time series residual suggests that the values are obtained in function of previous values, which is not showing independence between values [277]. The revision of autocorrelation can be done visually or using Durbin-Watson test or Ljung-Box Q-test. Considering a residual e_t given by the expression $e_t = \rho e_{t-1} + \nu_t$, both tests evaluate the following hypothesis:

Hypothesis $\mathbf{H_0}$: The data are independently distributed, which implies from previous expression of residual that $\rho = 0$ (i.e., the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

Hypothesis $\mathbf{H_1}$: The data are not independently distributed; they exhibit serial correlation, which implies alternative hypothesis $\rho \neq 0$.

- Durbin-Watson test [278, 279]:

  This test calculates for a residual e_t with number of observations n the following test statistic:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{4.59}$$

  Considering the sample autocorrelation of the residuals $\hat{\rho}$, the test statistic d is approximately equal to $2(1 - \hat{\rho})$. The value of d lies between 0 and 4, with $d = 2$ indicating no autocorrelation. There is evidence of positive serial correlation if $d < 2$. Therefore, small values of d indicate successive error

terms are positively correlated. If $d > 2$, successive error terms are negatively correlated, which can imply underestimation of the level of statistical significance α.

- Ljung-Box Q-test [246, 280]:

  This test evaluates the expression below considering a number of observations n:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{4.60}$$

  where $\hat{\rho_k}$ is the sample autocorrelation at lag k, and h is the number of tested lags. Under the null hypothesis the statistic Q asymptotically follows a chi-squared distribution with h degrees of freedom. The critical region for rejection of the hypothesis of randomness at significance level α corresponds to

$$Q > \chi^2_{1-\alpha,h} \tag{4.61}$$

  where $\chi^2_{1-\alpha,h}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with h degrees of freedom.

In an equivalent way, the presence of heteroscedasticity in the time series residual suggests that the variance alongside the values is not constant [246]. In order to develop this test, it is required to explain how Autoregressive Conditional Heteroskedasticity (ARCH) models are constructed. Assuming ϵ_t the series error term (represented by a stochastic component z_t) is a strong white noise process, and a time-dependent standard deviation σ_t, defined as $\epsilon_t = \sigma_t z_t$. The series σ_t^2 is modelled as follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\epsilon_{t-i}^2 \tag{4.62}$$

where $\alpha_0 > 0$ and $\alpha_i \geq 0$, $i > 0$. Heteroscedasticity is checked by using Ljung-Box Q-test from squared residuals or the Engle's ARCH test, which evaluate the following hypothesis:

Hypothesis H_0: The series of residuals σ_t exhibits no conditional heteroscedasticity (any ARCH effects).

Hypothesis H_1: An ARCH(q) model describes the series.

- Engle's ARCH test [246, 281]:

  The test statistic of the data with a sample size n using this approach corresponds to the Lagrange multiplier statistic nR^2, where R^2 corresponds to the coefficient of determination from fitting the ARCH(q) model for a number of lags q using regression. The asymptotic distribution of the test statistic under the null hypothesis is chi-square with q degrees of freedom.

- Ljung-Box Q-test from squared residuals: It is based on the same Ljung-Box Q-test statistic, but it is required to square the values of the analysed residuals. Rejecting the null hypothesis suggests presence of autoregressive conditional heteroscedasticity.

Showing all the tests that can be used to valuate distribution, they were applied over residuals to analyse the results obtained after predictions. Unless it is indicated something different, all the test applied in this thesis considered a significance level α=95%. For normality tests, only Anderson-Darling test, Lilliefors test and One-sample and One-sample Kolmogorov-Smirnov test are enough to evaluate the normality condition of residuals. Results are summarised in Tables J.31 and J.32 for the DMDc approach and Tables J.33 and J.34 for NARX approach, Appendix J.1. All of the tested methods failed on the normality test of the residuals, indicating that the results obtained so far can be further improved. Autocorrelation is a critical condition that needs to be checked in the residuals of a model. In the linear approach, both the training and validation datasets exhibited autocorrelation, as confirmed by the Durbin-Watson test. However, the coefficient d of around 2.4 indicates that the correlation observed is not critical. (Values of d above four indicate a critical condition [246]). To refine the models and improve their performance, it is necessary to evaluate the variables used in the regression. These variables include the input vector consisting of consumed/injected power, irradiance levels, and the proposed metrics M^P, M^Q, and the average normalised covariance. The measured vector corresponds to the voltage measurements. The next steps focus on processing and selecting the relevant inputs for the modelling approach in order to address the autocorrelation issue.

4.4.2 Data processing and selection

At this stage, the first obtained models can serve as reference for comparing and improving the modelling performance. However, the variables used as inputs in the modelling process have not been thoroughly analysed from a statistical perspective. The next step, as outlined in Algorithm 4.1, involves selecting relevant

data and analysing the variables used in the model. One approach to simplify the model is by analysing the collinearity between the explanatory variables and evaluating their impact by adding responses at different time lags. This analysis helps to explain the model's quasi-dynamics more concisely, which is crucial for predicting the voltage to be controlled. The current models should be examined, as the characteristics of the system may deviate from ideal conditions. The results presented so far have been obtained using the complete dataset of one thousand days. However, it is possible to reduce the dataset to focus only on critical values.

In this case, only the days in which any voltage variation exceeds $\pm 1.15\%$ are selected, resulting in a considerable reduction in the number of days to around 120 while still capturing voltages within the range of $\pm 3\%$. The remaining days represent the "most probable" scenario, where minimal prediction and control actions are required, and including them in the model would not be necessary. With this reduced dataset, the predicted voltages and the predictors used can be analysed to simplify the model complexity. When predicting the quasi-dynamics, correlated predictors can still be used without the need to separate their effects. However, it becomes problematic if the scenarios involve relationships between predictors and require a historical analysis of the contributions of various predictors. This situation is similar to multicollinearity, which occurs when two or more predictor variables in a multiple regression provide similar information [282–284]. Multicollinearity does not affect the predictive power of the model, but it can inflate the variance of individual predictor variables. Therefore, it is desirable to reduce this effect, particularly in the exogenous variables used in the model.

As the first approach, it is desired to check exogenous data that achieve critical values at lag 0, which variables are important or improve the observation of the phenomenon to be modelled. It is important to highlight that this is not a causality analysis, which is complex to give between variables in time-series. Causation is different from correlation, nor causation and forecasting [246, 260] It is desired to know if a variable x is useful to predict a variable y, but this is not saying that x is causing y. Only the data around values close to the critical scenarios will be considered in this case. It could be the presence of confounding (a variable that influences both predictor and response variable) that makes it difficult to determine if it is related to causation with others. However, it could not necessarily affect the prediction. Nevertheless, correlations are useful for predicting, even when there is a confounding or no causal relationship between the two variables. After reducing the dataset to only 120 days with 160 critical voltage values, it is performed a collinearity analysis with these remaining data. The idea is to check which variables are highly correlated in a regression model structure, which

reduces the precision of the estimated coefficients [260].

Figure 4.19 shows the results of computing correlation analysis of the remaining data. Here, the voltages on phase C at nodes 81, 84, and 85 are highly correlated, which is explained using the electric distance concept embedded into the covariance relationship between voltages, and explained in the previous chapter. For this purpose, the modelling can consider only one of these three nodes (in this case, the node showing higher voltage variation). The others two will follow the same response (assuming a radial topology, a common distribution system in most cases). The same applies for each phases on nodes 57 and 60, which are also highly correlated. Therefore, the voltages that remain after this the analysis are ΔV S85C,ΔV S81A, ΔV S81B, ΔV S60A, ΔV S60B and ΔV S60C.

It is performed the same analysis for the variables used as regressors. These consist of regressors used as potential signals in the control approach and other exogenous variables. The contribution of these potential control signals is useful in the model, and they are not required to reduce the number of delays. However, the exogenous variables must be processed since it is desired to reduce the variance. In this case, the input signal for controlling are the metrics M^P and M^Q, while the input signals are the average normalised covariance (for nodes 57, 60, 81, 84 and 85), consumed (nodes 84 and 85), power injections (node 85) and the solar irradiance.

Figure 4.20 shows the results of computing correlation analysis of these variables. It is shown that average normalised covariance on phase C at nodes 81, 84 and 85 are highly correlated, in a similar way as presented in previous case for voltage values. Same applies for each phase on nodes 57 and 60, which are also highly correlated. The power consumed in both nodes are not correlated, while the power injected due to renewable, and the irradiance level is also highly correlated. In this case, the irradiance is chosen as the variable to be used in the model.

Table 4.10: Belsley collinearity diagnosis for observed voltages

sValue	condIdx	ΔV S85C	ΔV S84C	ΔV S81C	ΔV S60C	ΔV S57C
2.0891	1.000	0.0000	0.0000	0.0000	0.0001	0.0001
0.7916	2.6390	0.0000	0.0000	0.0001	0.0014	0.0015
0.0861	24.2566	0.0147	0.0002	0.0416	0.0000	0.0023
0.0381	54.8112	0.0003	0.0000	0.0006	**0.9918**	**0.9940**
0.0078	268.7685	**0.9849**	**0.9998**	**0.9576**	0.0066	0.0021

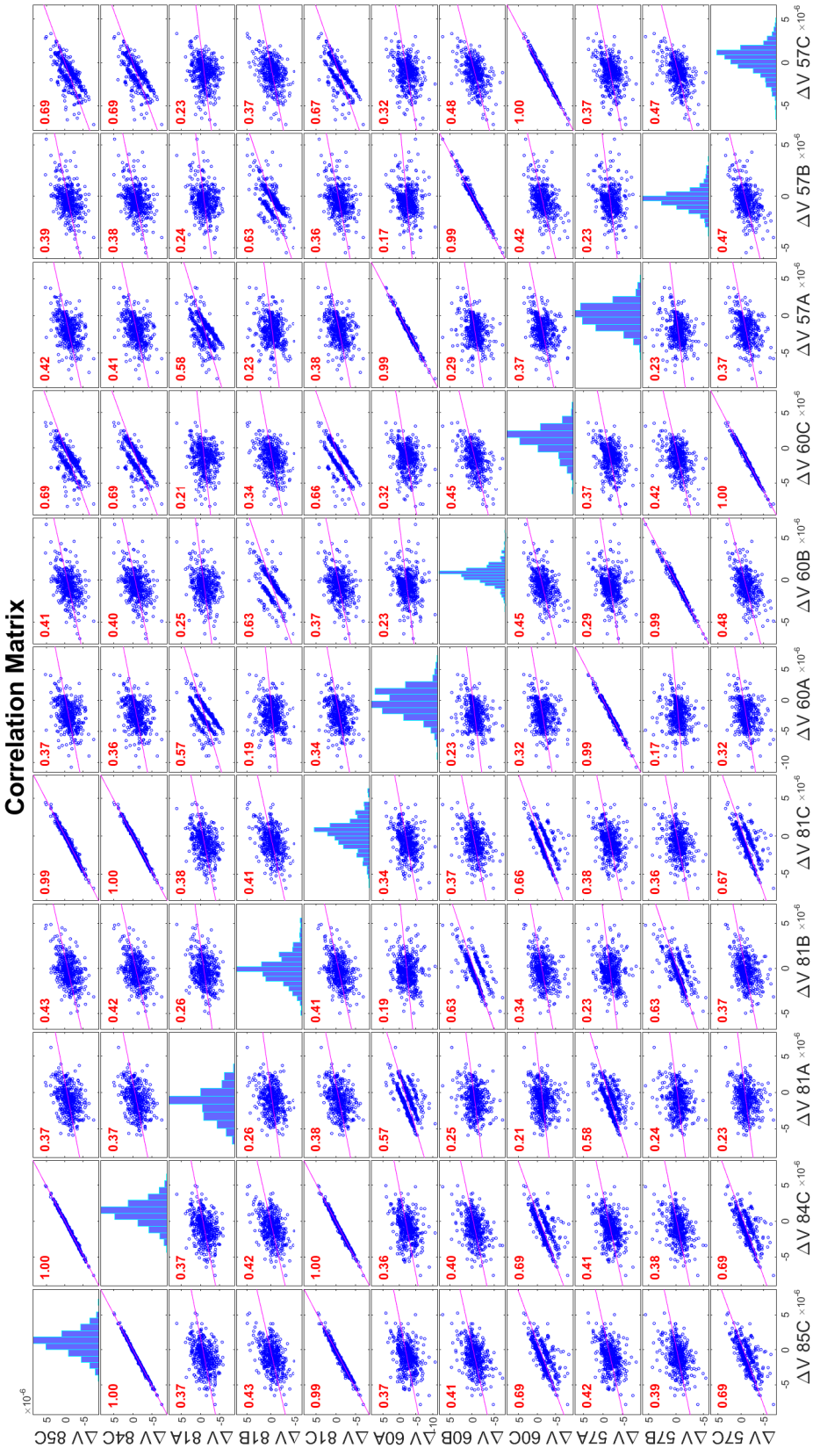


Figure 4.19: Histogram and correlations of observed voltages (only days for critical values)

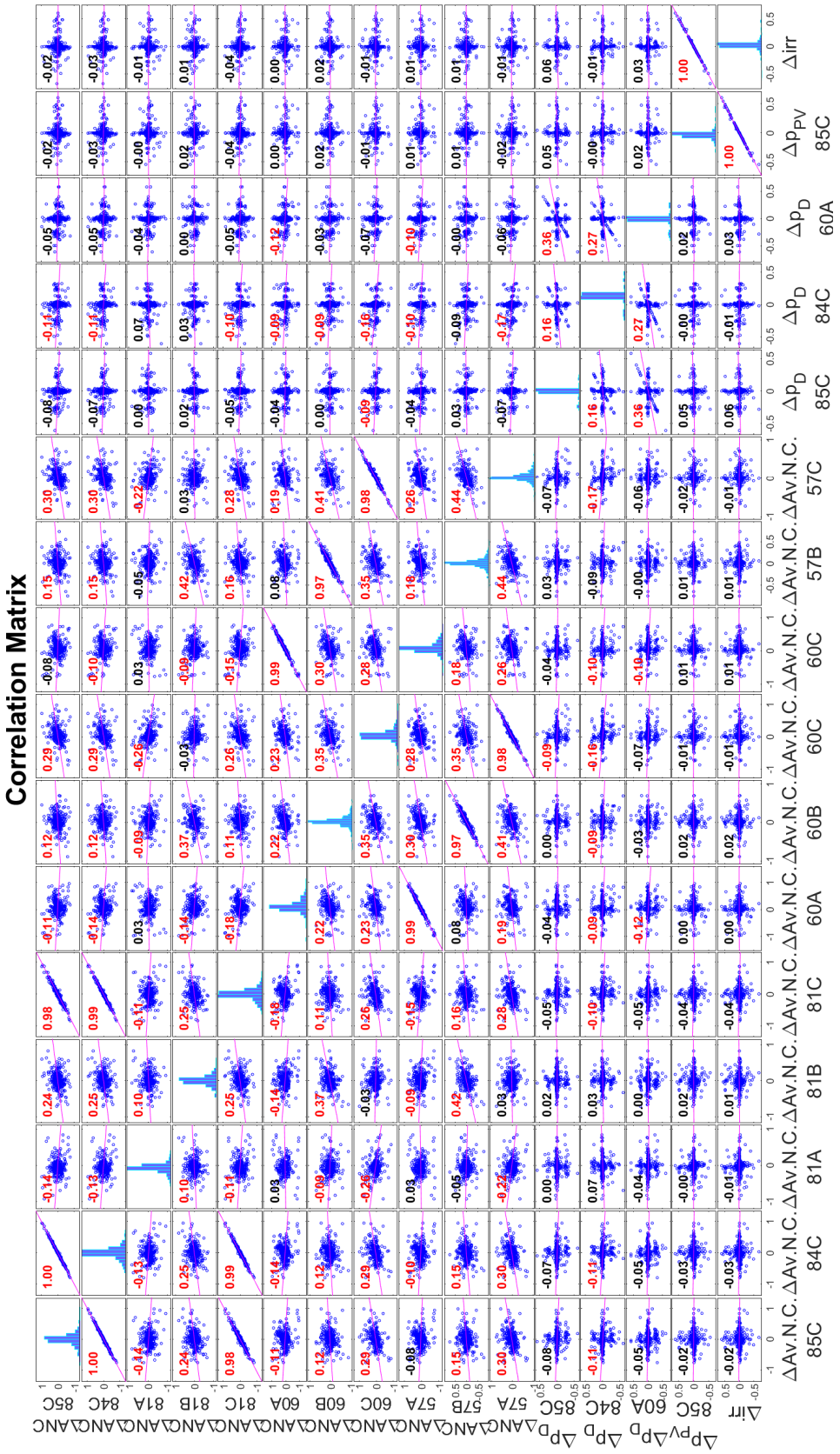


Figure 4.20: Histogram and correlations of exogenous inputs (only for critical values)

Table 4.11: Belsley collinearity diagnosis for exogenous regressors in input

sValue	condIdx	ΔAv. norm. cov. S85C	ΔAv. norm. cov. S84C	ΔAv. norm. cov. S81C	ΔAv. norm. cov. S60C	ΔAv. norm. cov. S57C	ΔP_D S85C	ΔP_D S84C	ΔP_G S85C	Irrad.
1.8313	1	0.0000	0.0000	0.0001	0.0003	0.0003	0.0012	0.0043	0.0000	0.0000
1.4214	1.2884	0.0000	0.0000	0.0000	0.0004	0.0004	0.0026	0.0015	0.0017	0.0017
1.3140	1.3937	0.0000	0.0000	0.0000	0.0018	0.0018	0.0362	0.0211	0.0002	0.0002
1.0514	1.7418	0.0000	0.0000	0.0000	0.0005	0.0006	0.3419	0.3570	0.0001	0.0001
0.8757	2.0914	0.0000	0.0000	0.0000	0.0000	0.0000	0.5526	0.5885	0.0001	0.0000
0.1329	13.7817	0.0089	0.0001	0.0259	0.0275	0.0367	0.0479	0.0038	0.0000	0.0001
0.0730	25.0712	0.0038	0.0000	0.0091	**0.9179**	**0.9284**	0.0047	0.0001	0.0003	0.0004
0.0645	28.3775	0.0000	0.0000	0.0000	0.0007	0.0009	0.0092	0.0002	**0.9968**	**0.9958**
0.0100	183.8411	**0.9872**	**0.9999**	**0.9649**	0.0510	0.0309	0.0038	0.0234	0.0008	0.0016

This is also confirmed using the Belsley collinearity diagnosis, which assess the strength of collinearity and possible sources among variables in a multiple linear regression structure, assuming that regressors are following a Gaussian distribution function [285, 286]. Only for illustration, results after applying this on phase C are shown in Tables 4.10 and 4.11. Values highlighted in both tables showed the variables that have a high strength of collinearity, which coincide with those obtained from the Figures 4.19 and 4.20. Therefore, the exogenous variables that remain after this the analysis are ΔAv. norm. cov. S81A, ΔAv. norm. cov. S81B, ΔAv. norm. cov. S81C, ΔAv. norm. cov. S57A, ΔAv. norm. cov. S60B, Δ Av. norm. cov. S60C, ΔP_D S85C, ΔP_D S84C and ΔP_{PV} S85C. The inputs potentially used for control remain the same on the same phase: M^P 60A-57A, M^Q 60A-57A, M^P 60B-57B, M^Q 60B-57B, M^P 60C-57C, M^Q 60C-57C, M^P 81C-84C, M^Q 81C-84C, M^P 84C-85C and M^Q 84C-85C.

It is also recommended to work with data that follows a normal distribution. This can be checked used the methods presented in the previous step for the residuals. If the data does not follow a normal distribution, it can be transformed using a Box-Cox transformation [260, 287], which is part of the power transform family function and is defined for positive and negative values as follows, respectively:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases} \tag{4.63a}$$

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(y_i + \lambda_2) & \text{if } \lambda_1 = 0, \end{cases} \tag{4.63b}$$

The first transformations hold for $y_i > 0$, while the second for $y_i > -\lambda_2$. The parameters λ, λ_1 and λ_2 are estimated using goodness-of-fit tests and profile likelihood functions.

Figures 4.19 and 4.20 show the histograms of the obtained voltages scenarios and the exogenous variables used in the input and the reduced dataset that represent the 120 days. Results after running normality test for the selected variables on phase C are presented in Tables J.35 and J.36. All the test mentioned in previous section were used to evaluate the distributions for all the variables (including the input used for controlling). All distribution showed to follow a Gaussian distribution according to the Cramer-Von Mises test, as it confirmed in the result of Rejecting H_1. Similar results can be obtained for phases A and B. Therefore, it can be concluded that it is not required to transform any of these analysed variable, and all can be used directly in the new model generation.

Finally, a selection of corresponding lags is done using a combination of cross-correlation analysis and Granger-causality analysis in the time-series data of regressors previously selected. The first tool is mainly static analysis (because it does not consider information from previous time steps) and measures similarity of two time-series as a function of the lag of one with respect to the other. Figures 4.21 to 4.26 show the results obtained after applying this procedure.

The way how the correlation is selected depends on the size of the series, which in this case is high and it would make hard the decision of selecting relevant lags. Therefore, it is a problem to define the threshold in the correlation obtained from the cross-correlation analysis, which is defined in terms of the inverse of the amount of data $1/\sqrt{n}$, and that would make to be relevant every correlation obtained in the analysis. In this work, it is proposed to use Granger-causality analysis to complement this analysis by providing a much more stringent criterion for causation than simply observing high correlation with some lag-lead relationship. Therefore, both static responses captured in the cross-correlation analysis and the dynamical response obtained from the Granger-causality analysis are contrasted for selecting the best lags [246, 260].

The Granger-causality analysis is an alternative to avoid thinking about causality in time-series analysis. This statistical hypothesis test determines whether one time series is useful in predicting another. An evolving-time variable $x(t)$ "Granger-causes" another variable $y(t)$ if predictions of $y(t)$ based on its own past values and on the past values of $x(t)$ are better than predictions of $y(t)$ based purely on its own past values, i.e., $x(t)$ helps to predict $y(t)$.

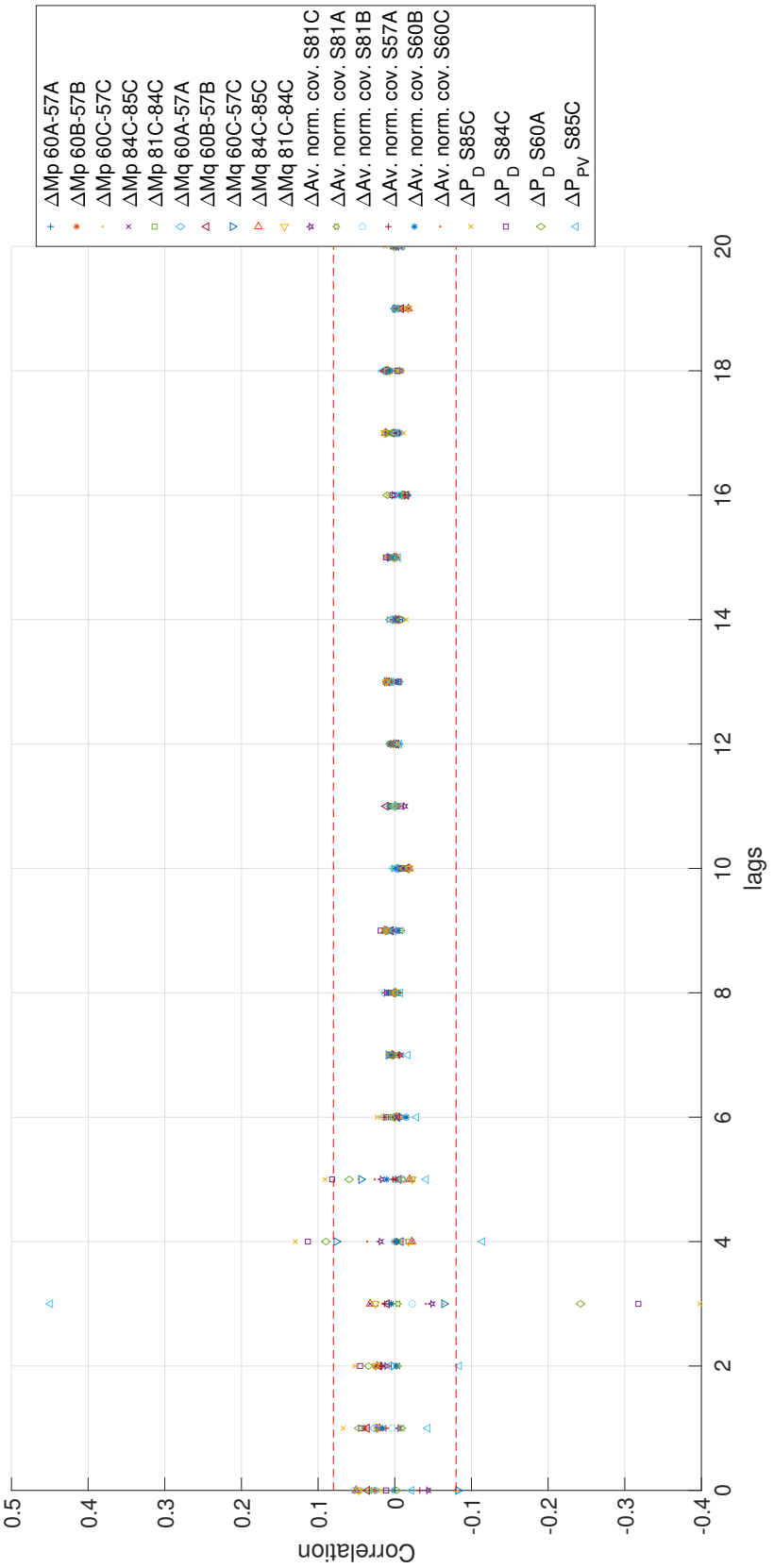


Figure 4.21: Cross-correlation analysis for first-lags relevant regressors at node S85C

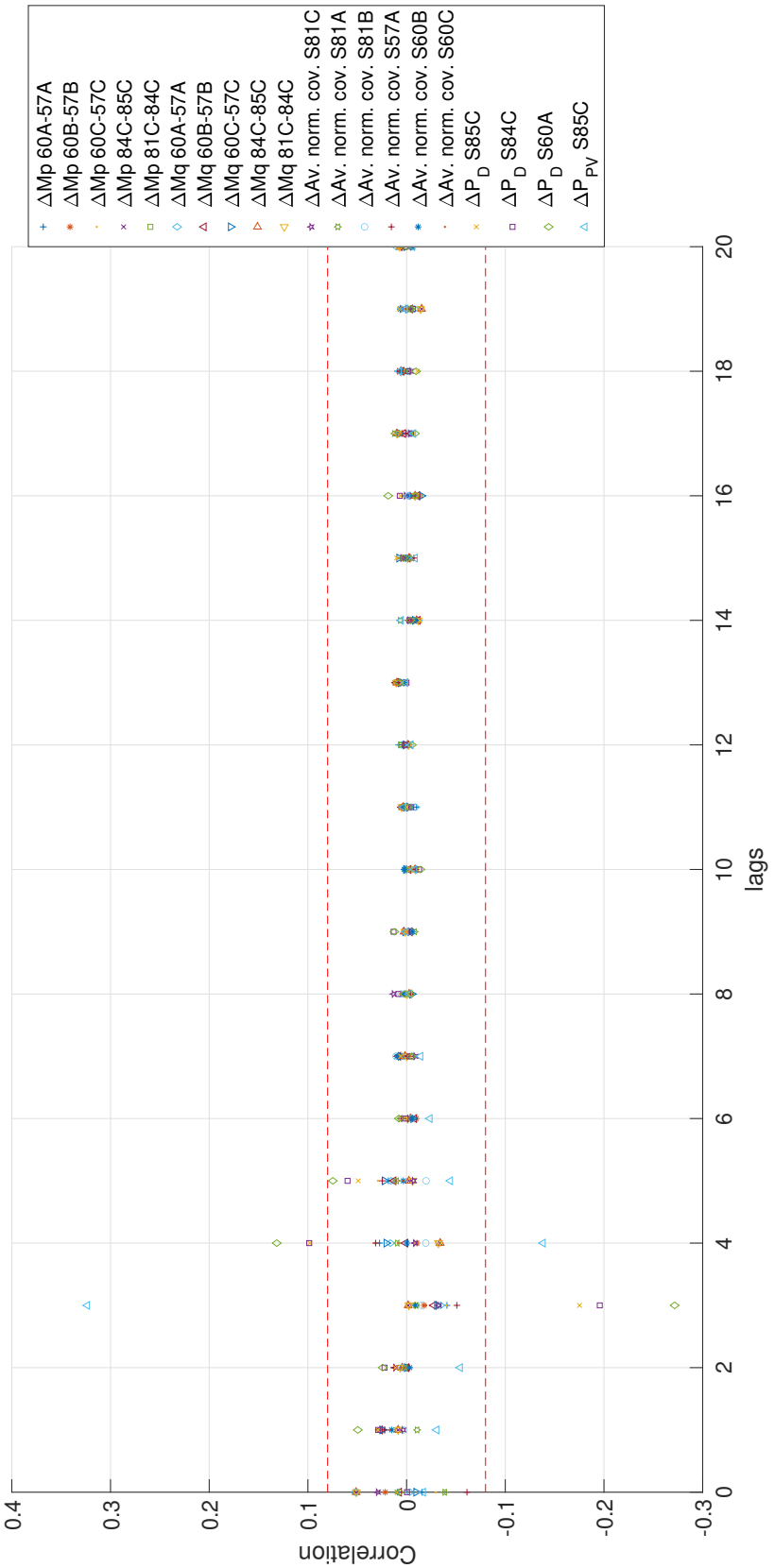


Figure 4.22: Cross-correlation analysis for first-lags relevant regressors at node S81A

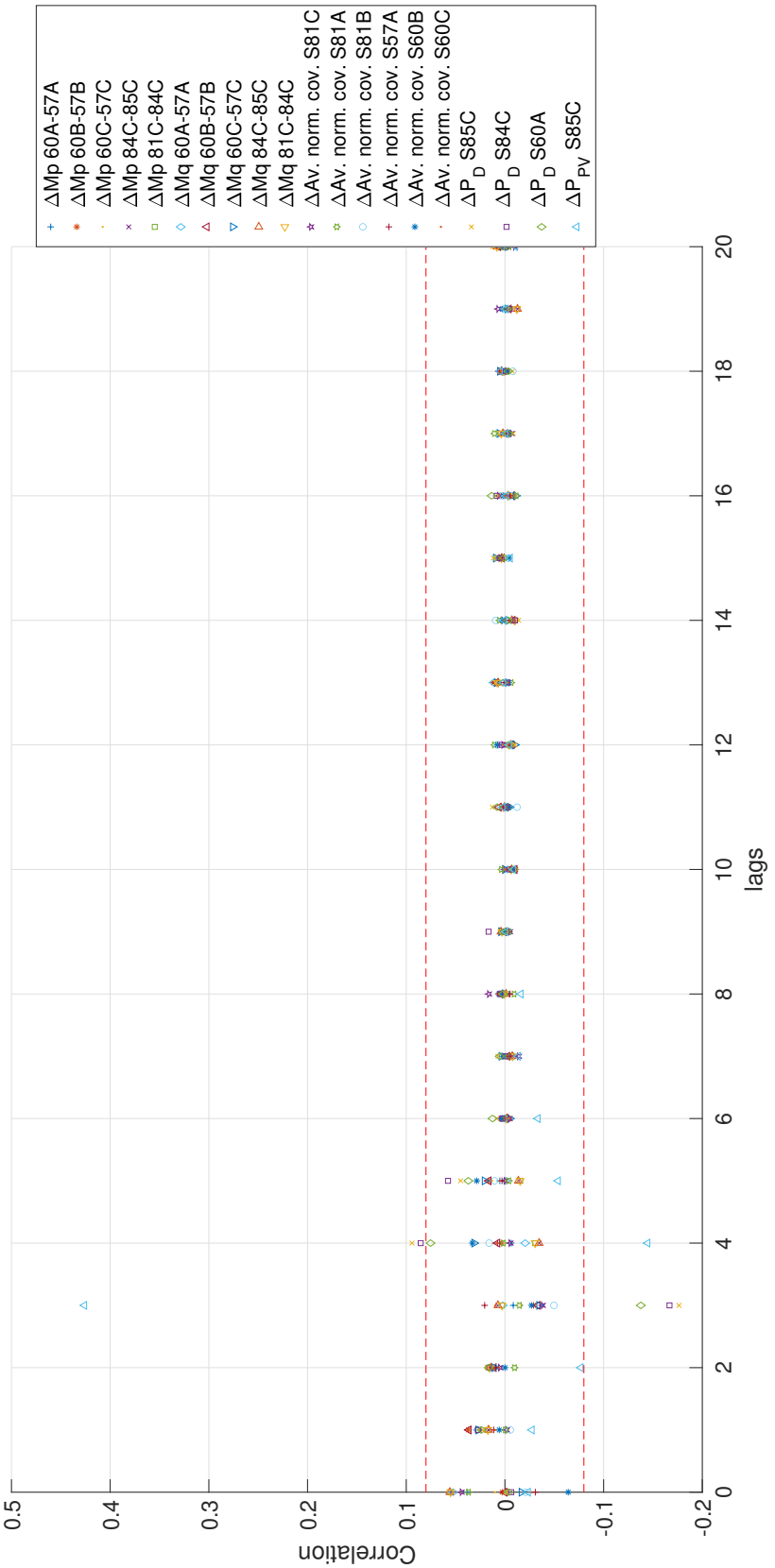


Figure 4.23: Cross-correlation analysis for first-lags relevant regressors at node S81B

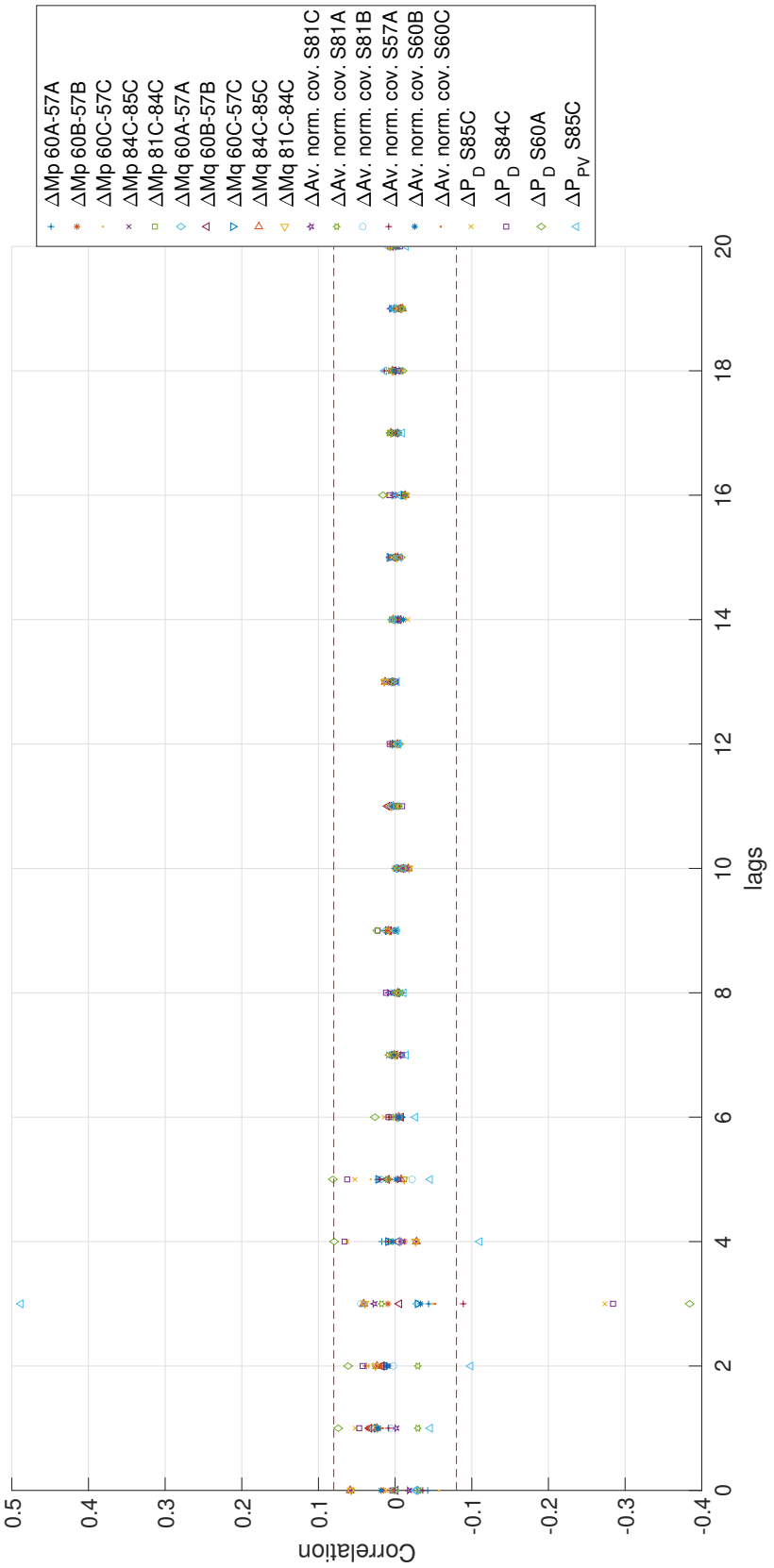


Figure 4.24: Cross-correlation analysis for first-lags relevant regressors at node S60A

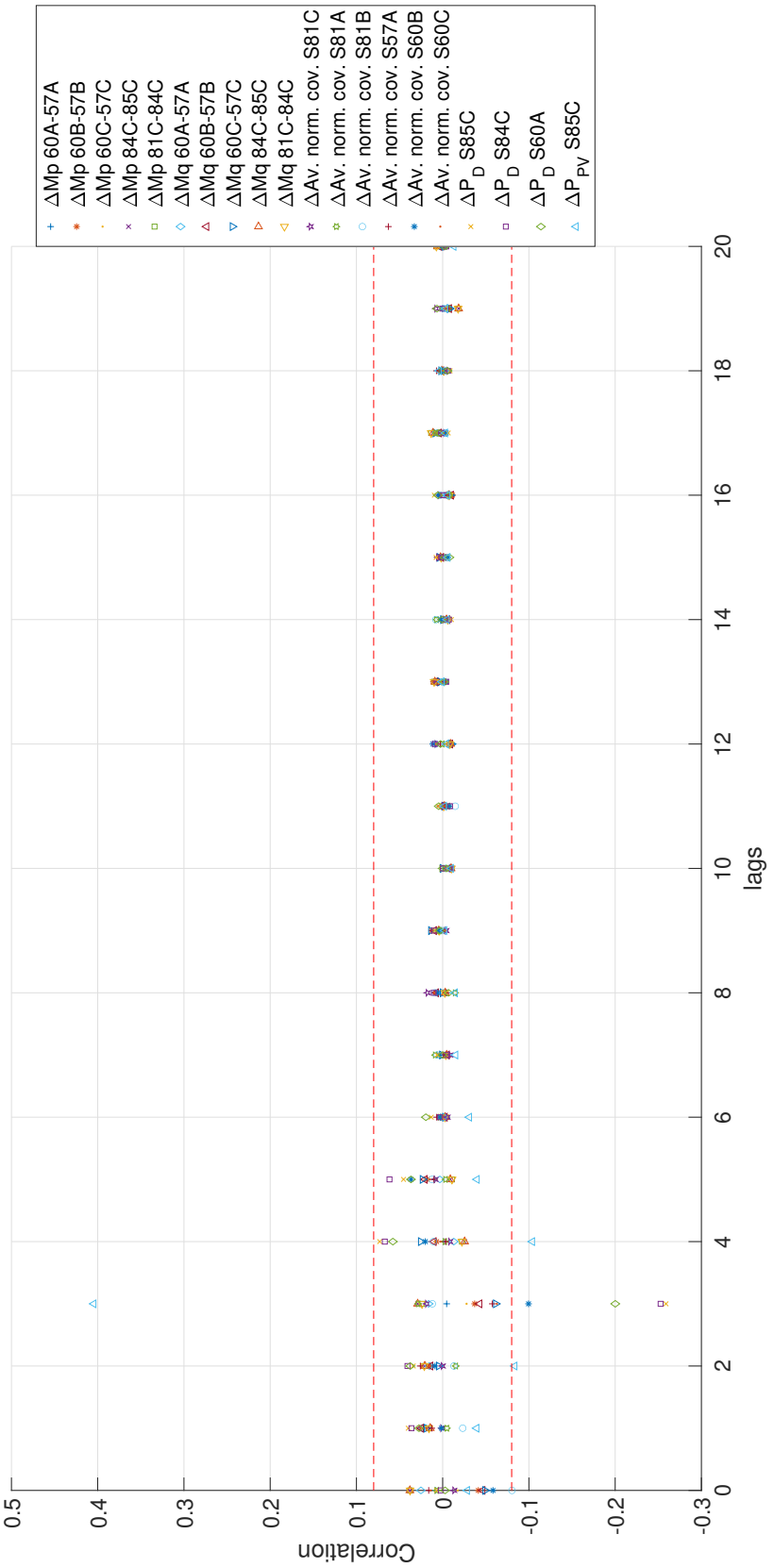


Figure 4.25: Cross-correlation analysis for first-lags relevant regressors at node S60B

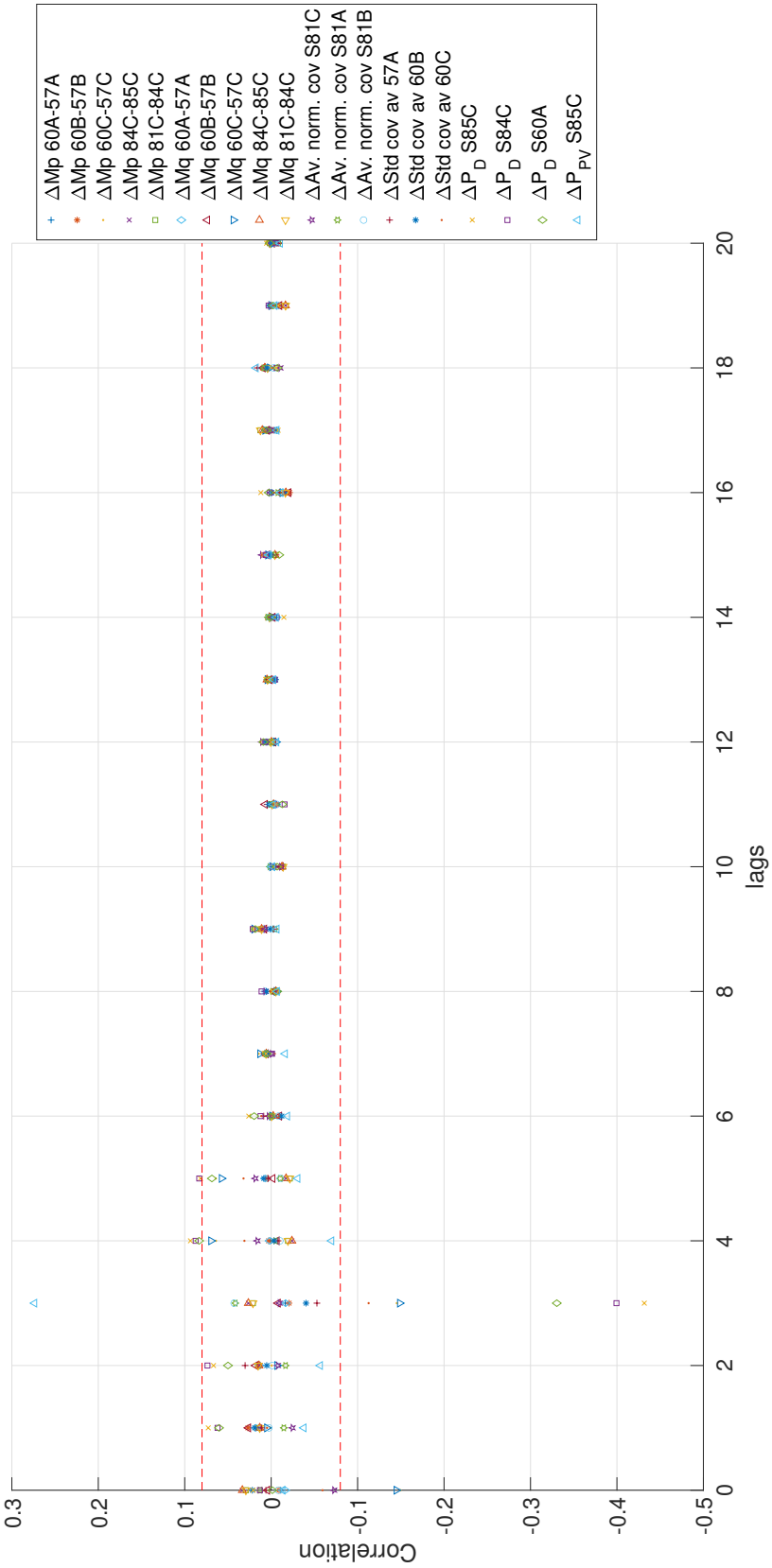


Figure 4.26: Cross-correlation analysis for first-lags relevant regressors at node S60C

This analysis is done by fitting two Vector Autoregressive Models (VARs) to the time series. A time series $x(t)$ is called a Granger cause of another time series $y(t)$, if at least one of the elements β_τ for $\tau = 1, \dots, q$ is significantly larger than zero (in absolute value). For this, the following two models are used:

$$y(t) = \gamma_0 + \sum_{\tau=1}^{q} \gamma_\tau y(t-\tau) + e(t) \tag{4.64a}$$

$$y(t) = \alpha_0 + \sum_{\tau=1}^{q} \alpha_\tau y(t-\tau) + \sum_{\tau=1}^{q} \beta_\tau x(t-\tau) + \varepsilon(t) \tag{4.64b}$$

where $e(t)$ and $\varepsilon(t)$ are white Gaussian random vectors. For the test statistic implies that the model in (4.64a) does not add information or provides a better model of $y(t)$, when comparing it to the model in (4.64b). Intuitively, the null hypothesis requires that $\forall\tau, \beta_\tau = 0$.

To run the test, it is assumed that the future values cannot inform past values, and the variables related with the cause only informs the variable related with the effect and no other variables are able to provide information (once at time). This test evaluates the following hypothesis:

Hypothesis H_0: A lagged x-value do not explain the variation in y, i.e., $x(t)$ does not Granger-cause $y(t)$.

Hypothesis H_1: A lagged x-value does explain the variation in y, i.e., $x(t)$ does Granger-cause $y(t)$.

It is used F-test to assess the alternative hypothesis considering both regressions with a significance level of α (also for this thesis, α=95%). Results after applying this test are shown in Table 4.12, in which relevant time lags are presented for the analysed variables with respect of the analysed voltage nodes.

Based on the results obtained from Figures 4.21 to 4.26 and Table 4.12, it can be concluded that the relevant lags for predicting voltage are 3, and 4. This means that considering the previous 30 and 40 minutes of data will help in predicting the voltage in the next 10-minute step. The selected regressors can be a combination of the results obtained from both analyses. To further reduce the complexity of the model, redundant signals can be eliminated using backward elimination [246]. This process helps to reduce the number of regressors used in the analysis.

In summary, after this analysis, the voltages observed at the 11 nodes can be reduced to only 6 nodes for all phases. The total of 26 possible regressors can be reduced to 18 relevant regressors, taking into account the different lags. This reduction has several advantages, including improving model stability and

Table 4.12: Relevant lags highlighted after applying Granger-causality analysis

	ΔV S85C	ΔV S81A	ΔV S81B	ΔV S60A	ΔV S60B	ΔV S60C
ΔM^P 60A-57A	5,6,7	—	—	6,7,8	6,7,8	0,3,5
ΔM^P 60B-57B	—	—	—	—	4,5,6	—
ΔM^P 60C-57C	0	—	—	—	2,5	—
ΔM^P 84C-85C	8,9	—	8,9	—	—	6,7,9
ΔM^P 81C-84C	6,8,9	—	7,8,9	—	—	7,8,9
ΔM^Q 60A-57A	7	—	—	4,8	—	6,8
ΔM^Q 60B-57B	—	—	3	—	5,6	—
ΔM^Q 60C-57C	—	—	7	—	2,5	3,5,6,7
ΔM^Q 84C-85C	8,9	—	8,9	—	—	6,7,9
ΔM^Q 81C-84C	6,8,9	—	7,8,9	—	—	7,8,9
ΔAv. norm. cov. S81C	—	—	—	—	—	—
ΔAv. norm. cov. S81A	—	—	—	—	—	—
ΔAv. norm. cov. S81B	—	—	—	—	0	—
ΔAv. norm. cov. S57A	—	—	—	3	—	—
ΔAv. norm. cov. S60B	—	—	—	—	3	—
ΔAv. norm. cov. S60C	—	—	—	—	—	3
ΔP_D S85C	3,4,5	3,4	3,4,8	3,5	3,6,7,8	3,4,5,6
ΔP_D S84C	2,3,4,5	3,4	3,4	3,8	2,3	3,5,8
ΔP_D S60A	3,4	2,3	3	4,5	3	3,4,5,8
ΔP_{PV} S85C	3,4	3,4	3,4,6	3,4	3,4	3,4,5,9

reducing complexity. The specific variables to be selected from this reduced set will depend on the desired approach for developing the modelling, whether it is a MISO or MIMO approach, which will be discussed in more detail in the next section.

4.4.3 Creation of LTI model using revised data

Once relevant regressors and lags are selected, the process of linear regression is done again to obtain a model that helps on the prediction by detecting the voltage quasi-dynamics. In this step, which corresponds to Step 3 of Algorithm 4.1, the process of linear regression is performed again with the selected relevant regressors and common lags. This reduces the number of system inputs and outputs.

Table 4.13 presents the obtained regressors from the previous step that will be used in each approach, with common regressors highlighted in italics. Two structures are evaluated this time: a MISO system considering each output independently and only the relevant lags found for each output, and a MIMO system considering only the common relevant regressors and lags. Since the number of regressors is reduced, the MIMO model is developed for all phases together.

MISO						MIMO	
ΔV S85C	ΔV S81A	ΔV S81B	ΔV S60A	ΔV S60B	ΔV S60C	All nodes	
ΔM^P 60C-57C lag 0	ΔP^D S85C lag 3	ΔP^D S85C lag 4	ΔAv. norm. cov. S57A lag 3	ΔAv. norm. cov. S81B lag 0	ΔM^P 60C-57C lag 0	ΔP^D *S84C lag 3*	
ΔP^D S85C lag 3	ΔP^D S85C lag 4	ΔP^D *S84C lag 3*	ΔP^D S85C lag 3	ΔAv. norm. cov. S60B lag 3	ΔM^P 60C-57C lag 3	—	
ΔP^D S85C lag 4	ΔP^D *S84C lag 3*	ΔP^D S84C lag 4	ΔP^D *S84C lag 3*	ΔP^D S85C lag 3	ΔM^Q 60C-57C lag 3	—	
ΔP^D S85C lag 5	ΔP^D S84C lag 4	ΔP^D S60A lag 3	ΔP^D S60A lag 5	ΔP^D *S84C lag 3*	ΔAv. norm. cov. S60C lag 3	—	
ΔP^D *S84C lag 3*	ΔP^D S60A lag 3	ΔP^{PV} S85C lag 3	ΔP^{PV} S85C lag 3	ΔP^D S60A lag 3	ΔP^D S85C lag 3	—	
ΔP^D S84C lag 4	ΔP^{PV} S85C lag 4	ΔP^{PV} S85C lag 4	ΔP^{PV} S85C lag 4	ΔP^{PV} S85C lag 3	ΔP^D S85C lag 4	—	
ΔP^D S84C lag 5	—	—	—	ΔP^{PV} S85C lag 4	ΔP^D S85C lag 5	—	
ΔP^D S60A lag 3	—	—	—	—	ΔP^D *S84C lag 3*	—	
ΔP^D S60A lag 4	—	—	—	—	ΔP^D S60A lag 3	—	
ΔP^{PV} S85C lag 3	—	—	—	—	ΔP^D S60A lag 4	—	
ΔP^{PV} S85C lag 4	—	—	—	—	ΔP^{PV} S85C lag 3	—	

Table 4.13: Input selected for each approach

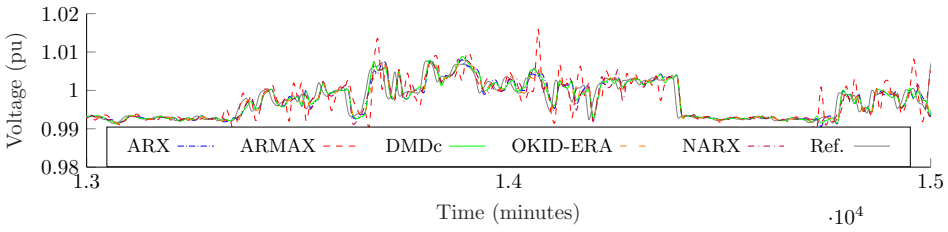
Since the effect of lags was evaluated to simplify the system model and improve prediction, both approaches consider the lagged inputs as independent regressors, which will impact the order of the models. Therefore, the autoregressive structures ARX, ARMAX, and NARX consider output delays up to lag 0 (as it is already considered in the input vector), and internal input delays up to lag 0. In an equivalent way to before, the first matrix $X^{'}$ for DMDc considers the matrix order equivalent to the ARX state-space representation in the exploration process. The OKID-ERA approach showed the best performance when the number of Markov parameters was set to 50. Similar to the first attempt at regression, all models were developed under similar conditions of dimension size to ensure that all models were comparable regarding their ability to represent the system dynamics using available data. It is not required for this thesis to obtain the optimal model for each algorithm, and therefore, a wide search space for model dimensions.

Training of all models was done using the portion of data that represents the critical cases presented before (critical 120 days). The validation of results was performed with the original dataset to make the results comparable with the first guess and detect any differences.

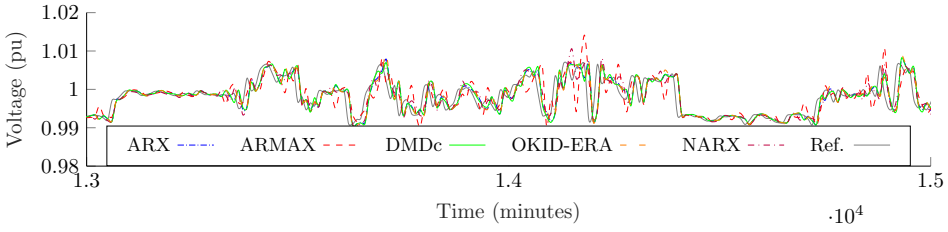
A portion of the results obtained from the MISO structure and the predicted voltages are presented in Figures 4.27 and 4.28 from the original training and validation dataset, respectively. Tables from 4.14 to 4.19 summarise general features of the models obtained on each case. Table 4.20 shows the performance for training on each node; Table 4.21 shows the performance for validation on each node. In a similar way, a portion of the results obtained from the MIMO structure and the predicted voltages are presented in Figures 4.29 and 4.30 from the original training and validation dataset, respectively; Table 4.22 summarises general characteristics of the models obtained. Table 4.23 and 4.24 show the performance for training and validation.

Table 4.14: Obtained models dimensions for voltage prediction at node S85C using selected regressors training dataset in MISO structure

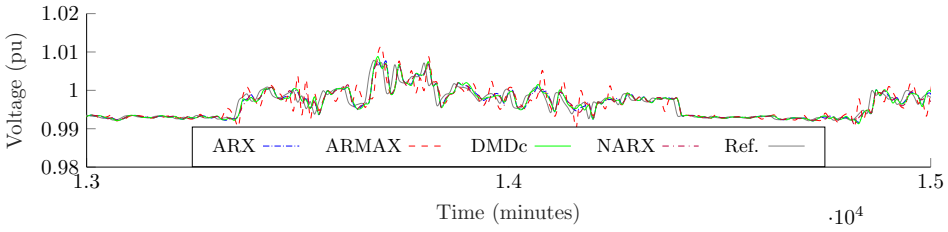
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension *A*	1x1	1x1	1x1	6x6	—
Dimension *B*	1x11	1x11	1x11	6x11	—
Dimension *C*	1x1	1x1	1x1	1x6	—
Dimension *D*	1x11	1x11	1x11	1x11	—
Comp. time (s)	4.33	10.00	0.37	3.84	8.86



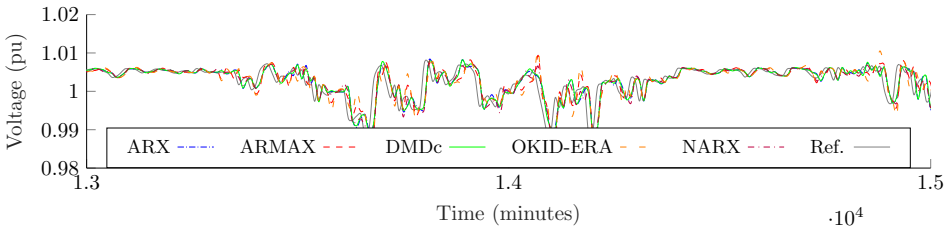
(a) Node S85C



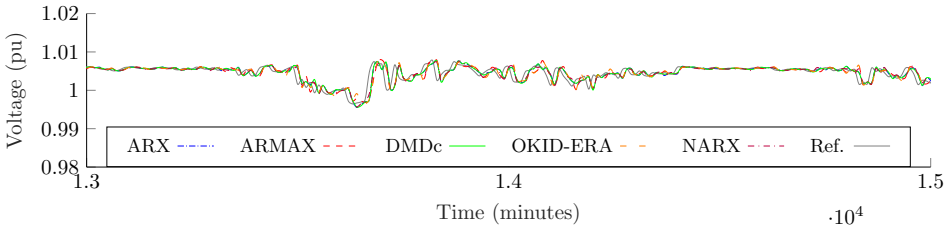
(b) Node S81A



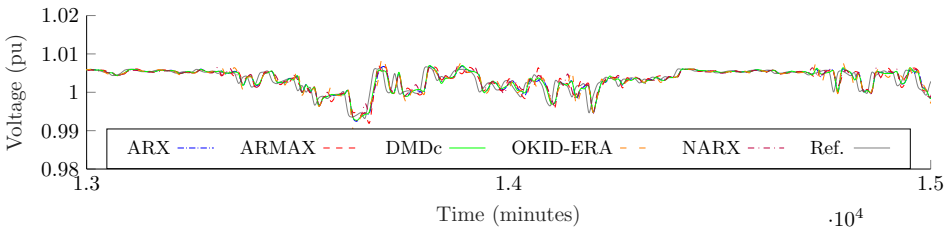
(c) Node S81B



(d) Node S60A

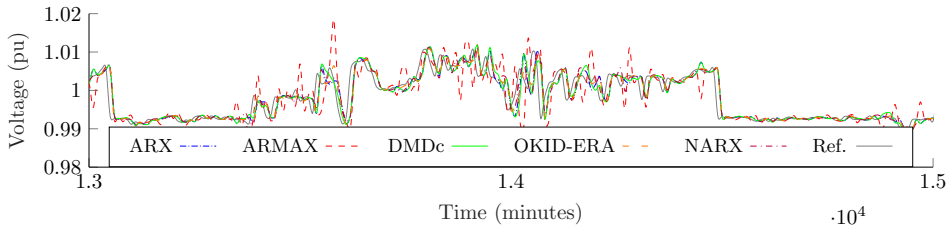


(e) Node S60B

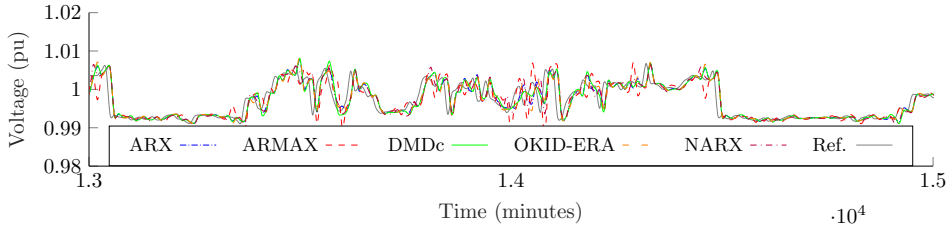


(f) Node S60C

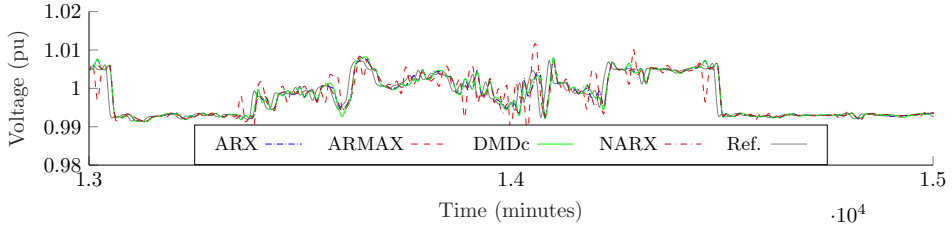
Figure 4.27: Portion of voltage predictions 1 step ahead using selected regressors training dataset in MISO structure



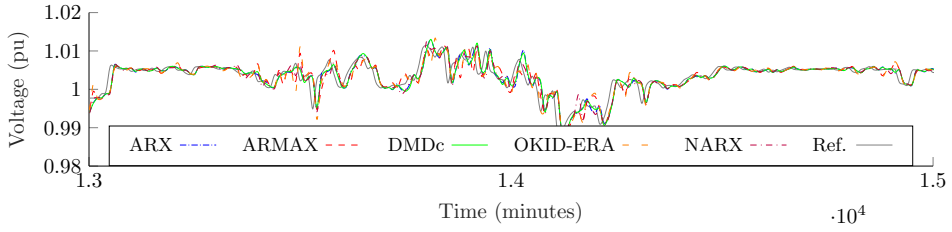
(a) Node S85C



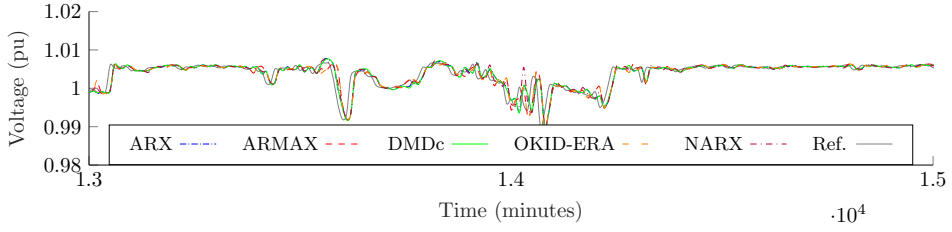
(b) Node S81A



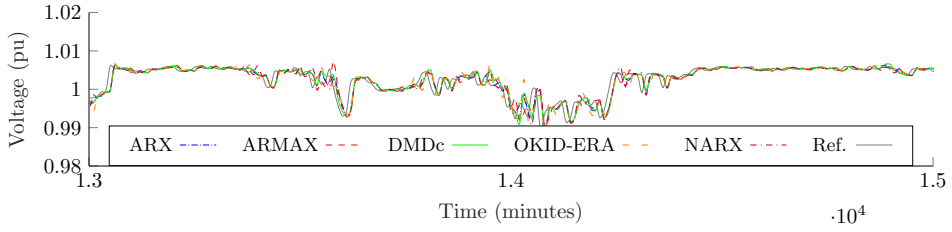
(c) Node S81B



(d) Node S60A



(e) Node S60B



(f) Node S60C

Figure 4.28: Portion of voltage predictions 1 step ahead using selected regressors validation dataset in MISO structure

Table 4.15: Obtained models dimensions for voltage prediction at node S81A using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension *A*	1x1	1x1	1x1	6x6	—
Dimension *B*	1x6	1x6	1x6	6x6	—
Dimension *C*	1x1	1x1	1x1	1x6	—
Dimension *D*	1x6	1x6	1x6	1x6	—
Comp. time (s)	0.38	2.05	0.09	2.70	1.90

Table 4.16: Obtained models dimensions for voltage prediction at node S81B using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension *A*	1x1	1x1	1x1	6x6	—
Dimension *B*	1x6	1x6	1x6	6x6	—
Dimension *C*	1x1	1x1	1x1	1x6	—
Dimension *D*	1x6	1x6	1x6	1x6	—
Comp. time (s)	0.14	1.56	0.07	2.72	1.58

Table 4.17: Obtained models dimensions for voltage prediction at node S60A using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension *A*	1x1	1x1	1x1	6x6	—
Dimension *B*	1x6	1x6	1x6	6x6	—
Dimension *C*	1x1	1x1	1x1	1x6	—
Dimension *D*	1x6	1x6	1x6	1x6	—
Comp. time (s)	0.14	1.41	0.07	2.75	1.74

Table 4.18: Obtained models dimensions for voltage prediction at node S60B using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension *A*	1x1	1x1	1x1	6x6	—
Dimension *B*	1x7	1x7	1x7	6x7	—
Dimension *C*	1x1	1x1	1x1	1x6	—
Dimension *D*	1x7	1x7	1x7	1x7	—
Comp. time (s)	0.14	1.70	0.07	3.09	2.16

Table 4.19: Obtained models dimensions for voltage prediction at node S60C using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension A	1x1	1x1	1x1	6x6	—
Dimension B	1x11	1x11	1x11	6x11	—
Dimension C	1x1	1x1	1x1	1x6	—
Dimension D	1x11	1x11	1x11	1x11	—
Comp. time (s)	0.18	1.72	0.07	3.73	1.70

Table 4.20: Results of models for voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.64	0.26	0.63	0.66	0.63
R^2 S81A	0.46	0.28	0.46	0.45	0.44
R^2 S82B	0.67	0.31	0.67	-3.78	0.66
R^2 S60A	0.59	0.52	0.60	0.49	0.58
R^2 S60B	0.70	0.52	0.70	0.67	0.70
R^2 S60C	0.68	0.67	0.68	0.59	0.67
NRMSE S85C	0.11	0.16	0.11	0.11	0.11
NRMSE S81A	0.14	0.17	0.14	0.15	0.15
NRMSE S82B	0.13	0.19	0.13	0.49	0.13
NRMSE S60A	0.08	0.09	0.08	0.09	0.08
NRMSE S60B	0.05	0.09	0.05	0.06	0.06
NRMSE S60C	0.06	0.06	0.06	0.07	0.06
AIC S85C	-3.65e5	-3.42e5	-3.64e5	-3.66e5	-3.64e5
BIC S85C	-3.64e5	-3.42e5	-3.64e5	-3.66e5	-3.64e5
AIC S81A	-3.64e5	-3.55e5	-3.64e5	-3.63e5	-3.63e5
BIC S81A	-3.64e5	-3.55e5	-3.64e5	-3.63e5	-3.63e5
AIC S81B	-3.77e5	-3.54e5	-3.77e5	-2.94e5	-3.77e5
BIC S81B	-3.77e5	-3.54e5	-3.77e5	-2.94e5	-3.77e5
AIC S60A	-3.65e5	-3.61e5	-3.66e5	-3.58e5	-3.65e5
BIC S60A	-3.65e5	-3.61e5	-3.66e5	-3.58e5	-3.65e5
AIC S60B	-3.95e5	-3.61e5	-3.95e5	-3.91e5	-3.94e5
BIC S60B	-3.95e5	-3.61e5	-3.95e5	-3.91e5	-3.94e5
AIC S60C	-3.83e5	-3.92e5	-3.83e5	-3.75e5	-3.82e5
BIC S60C	-3.83e5	-3.92e5	-3.83e5	-3.75e5	-3.82e5

Table 4.21: Results of models for voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.63	0.24	0.62	0.65	0.62
R^2 S81A	0.45	0.27	0.45	0.44	0.43
R^2 S82B	0.65	0.28	0.65	-4.05	0.64
R^2 S60A	0.57	0.50	0.58	0.47	0.57
R^2 S60B	0.69	0.50	0.69	0.65	0.68
R^2 S60C	0.67	0.66	0.67	0.58	0.66
NRMSE S85C	0.11	0.16	0.11	0.11	0.11
NRMSE S81A	0.14	0.16	0.14	0.14	0.14
NRMSE S82B	0.13	0.19	0.13	0.50	0.13
NRMSE S60A	0.10	0.10	0.10	0.11	0.10
NRMSE S60B	0.06	0.10	0.06	0.06	0.06
NRMSE S60C	0.08	0.06	0.08	0.09	0.08
AIC S85C	-3.64e5	-3.41e5	-3.64e5	-3.66e5	-3.63e5
BIC S85C	-3.64e5	-3.41e5	-3.64e5	-3.65e5	-3.63e5
AIC S81A	-3.64e5	-3.55e5	-3.64e5	-3.63e5	-3.63e5
BIC S81A	-3.64e5	-3.55e5	-3.64e5	-3.63e5	-3.63e5
AIC S81B	-3.76e5	-3.53e5	-3.76e5	-2.92e5	-3.75e5
BIC S81B	-3.76e5	-3.53e5	-3.76e5	-2.92e5	-3.75e5
AIC S60A	-3.65e5	-3.60e5	-3.66e5	-3.58e5	-3.64e5
BIC S60A	-3.65e5	-3.60e5	-3.65e5	-3.58e5	-3.64e5
AIC S60B	-3.93e5	-3.60e5	-3.93e5	-3.90e5	-3.93e5
BIC S60B	-3.93e5	-3.60e5	-3.93e5	-3.90e5	-3.93e5
AIC S60C	-3.83e5	-3.91e5	-3.83e5	-3.75e5	-3.81e5
BIC S60C	-3.83e5	-3.91e5	-3.83e5	-3.75e5	-3.81e5

Table 4.22: Obtained models dimensions for voltage prediction using selected regressors training dataset in MIMO structure

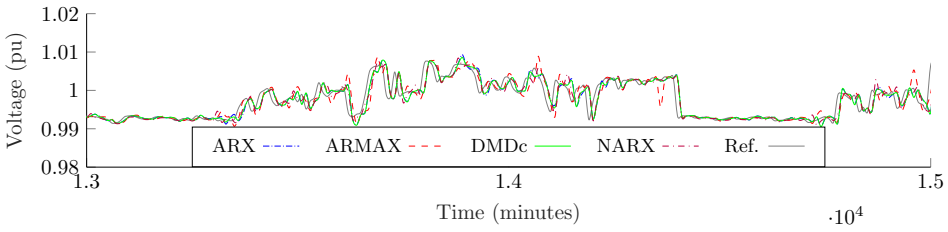
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
Dimension A	6x6	6x6	6x6	6x6	—
Dimension B	6x1	6x1	6x1	6x1	—
Dimension C	6x6	6x6	6x6	6x6	—
Dimension D	6x1	6x1	6x1	6x1	—
Comp. time (s)	2.52	5.30	0.16	2.85	5.35

Table 4.23: Results of models for voltage prediction using selected regressors training dataset in MIMO structure

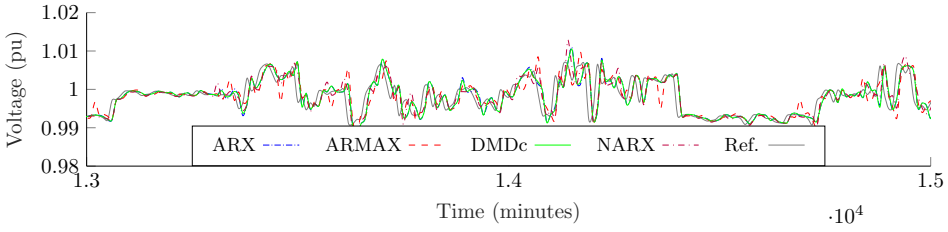
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.63	0.57	0.63	-5.60	0.63
R^2 S81A	0.42	0.35	0.43	-21.87	0.43
R^2 S82B	0.65	0.57	0.66	-1.48	0.65
R^2 S60A	0.59	0.50	0.59	-25.27	0.59
R^2 S60B	0.70	0.64	0.70	-6.34	0.69
R^2 S60C	0.68	0.65	0.68	-9.91	0.68
NRMSE S85C	0.11	0.12	0.11	0.48	0.11
NRMSE S81A	0.15	0.16	0.15	0.94	0.15
NRMSE S82B	0.13	0.15	0.13	0.35	0.13
NRMSE S60A	0.08	0.09	0.08	0.66	0.08
NRMSE S60B	0.06	0.06	0.06	0.27	0.06
NRMSE S60C	0.06	0.06	0.06	0.34	0.06
AIC	-3.74e5	-3.69e5	-3.74e5	-2.73e5	-3.74e5
BIC	-3.74e5	-3.69e5	-3.74e5	-2.73e5	-3.74e5

Table 4.24: Results of models for voltage prediction using selected regressors validation dataset in MIMO structure

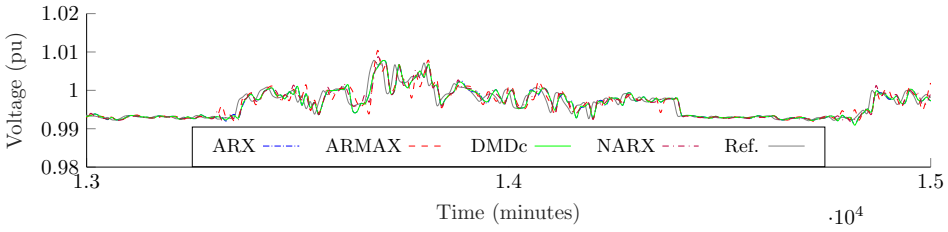
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.62	0.56	0.63	-6.07	0.62
R^2 S81A	0.42	0.33	0.42	-23.89	0.41
R^2 S82B	0.63	0.54	0.64	-1.66	0.63
R^2 S60A	0.58	0.48	0.58	-27.68	0.57
R^2 S60B	0.68	0.62	0.68	-6.85	0.68
R^2 S60C	0.67	0.64	0.67	-10.73	0.67
NRMSE S85C	0.11	0.12	0.11	0.49	0.11
NRMSE S81A	0.14	0.15	0.14	0.95	0.15
NRMSE S82B	0.14	0.15	0.13	0.36	0.14
NRMSE S60A	0.10	0.11	0.10	0.79	0.10
NRMSE S60B	0.06	0.07	0.06	0.30	0.06
NRMSE S60C	0.08	0.08	0.08	0.47	0.08
AIC	-3.73e5	-3.68e5	-3.74e5	-2.71e5	-3.73e5
BIC	-3.73e5	-3.68e5	-3.74e5	-2.71e5	-3.73e5



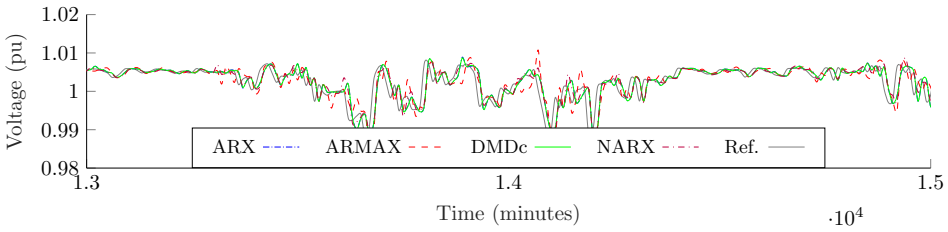
(a) Node S85C



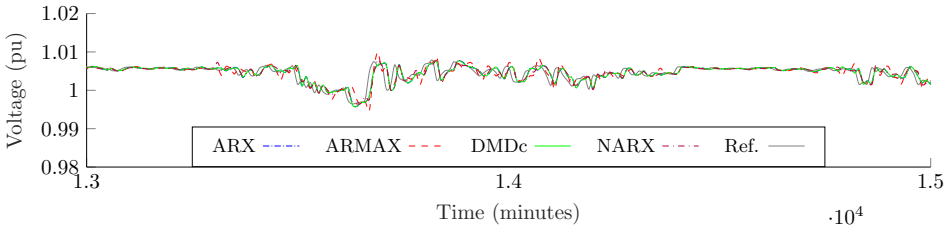
(b) Node S81A



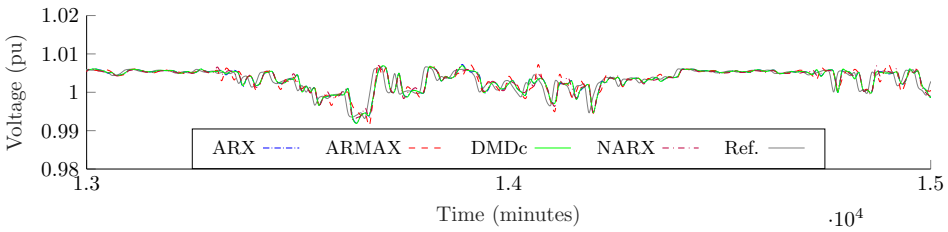
(c) Node S81B



(d) Node S60A

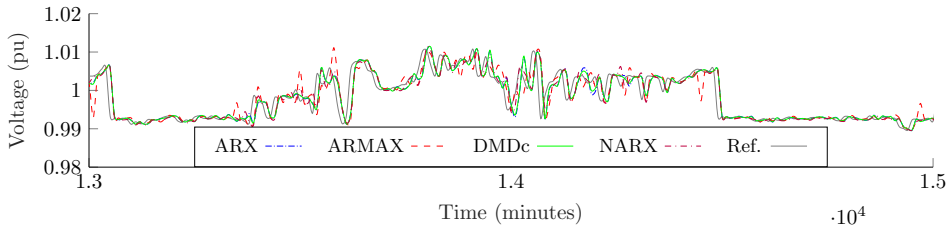


(e) Node S60B

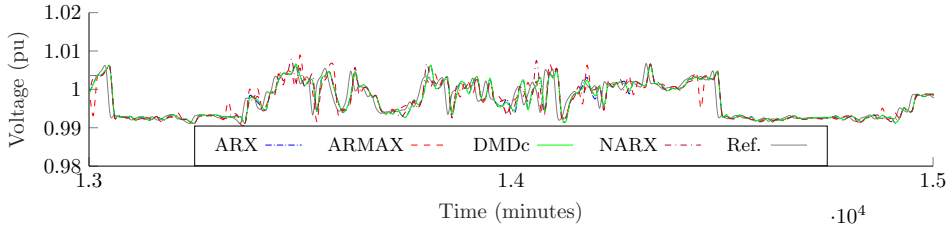


(f) Node S60C

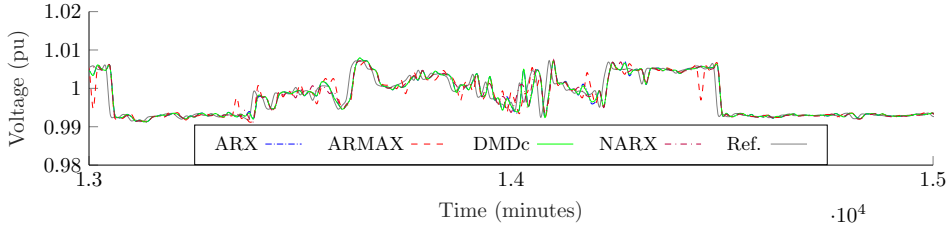
Figure 4.29: Portion of voltage predictions 1 step ahead using selected regressors training dataset in MIMO structure



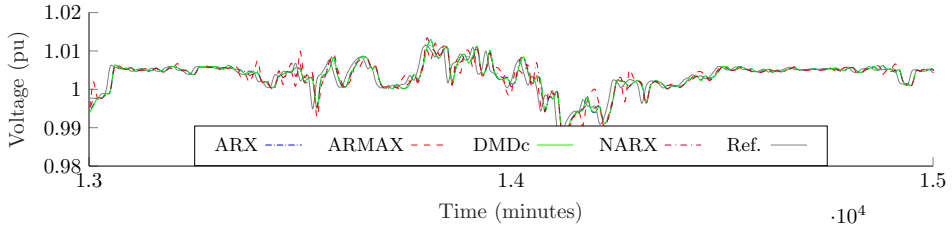
(a) Node S85C



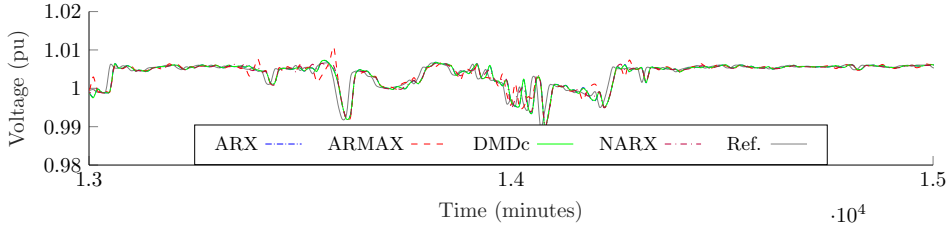
(b) Node S81A



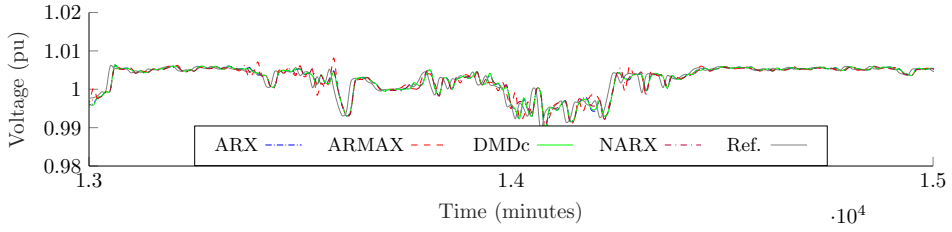
(c) Node S81B



(d) Node S60A



(e) Node S60B



(f) Node S60C

Figure 4.30: Portion of voltage predictions 1 step ahead using selected regressors validation dataset in MIMO structure

From the results obtained, it is shown that all models improved with the selected regressors. The results obtained from the predicted voltages in comparison with the first attempt presented from Figures 4.11 to 4.14 showed visually that all models responds in a similar way.

It is also confirmed from the similar values obtained for R^2 and NRMSE that were introduced in Tables 4.4, 4.5 and Tables 4.6 and 4.7. Even if the response from OKID-ERA were not satisfactory, the overall responses in both cases improved considerably. Computational time and model order were reduced in comparison with the first attempt, which is a good indicator for real-time applications.

It is consistently shown that the best response from all linear approaches considering complexity, computational time, AIC and BIC is DMDc (as mentioned in Section 4.4.1.2, AIC and BIC are presented as indicators of relative performance only for each compared model approach). NARX structure was not significantly better in comparison with the rest of procedures. In general, the R-squared index was not highly improved (most of cases values were equal or slightly reduced).

These results are satisfactory, considering all other features that were improved, without losing prediction capability. Nevertheless, a high value of R^2 is not always a good indicator of better performance[246], and there could be several reasons that produce a high value of R^2, such as data overfitting. To get the full picture, it must be considered R-squared values in combination with residual plots, in-depth knowledge of the subject area and other statistics.

Based on dimension size, computational time and performance metrics, the structure that showed better performance in overall was MISO. Nevertheless, it can be noticed that the number of inputs is higher in comparison with MIMO case. This helps on having more elements to describe the system, presuming that each node should be considered individually. With the idea of producing a model that presumes interaction between the nodes, MIMO approach can be more useful.

4.4.4 Checking validity of assumptions

Step 4 of Algorithm 4.1 corresponds to see if there is any improvement in the residuals obtained, to validate the assumptions done for the linear model approaches. Ideally, the obtained residual should represent white noise (normal distribution, no autocorrelation, and no heteroscedasticity). Therefore, it is required to see the characteristics of the residuals obtained after the regressor analysis.

Figures 4.31 and 4.32 presents the histogram, Q-Q plot and ACF components of the residuals for training and validation of DMDc (MISO approach), which showed the best performance in overall from the linear modelling approaches. Figures 4.33 and 4.34 presents the results for the residuals of NARX to compare

the performance. Figures 4.35 and 4.36 presents the histogram, Q-Q plot and ACF components of the residuals for training and validation of DMDc for MIMO approach. Figures 4.37 and 4.38 presents the results for the residuals of NARX.

From these figures, it is shown that for both MISO and MIMO approaches that, even if the residuals obtained are not following a normal distribution, the heavy-tailed shape is improved (compared with the corresponding residuals on phase C in Figures 4.15 and 4.16 for DMDc approach and Figures 4.17 and 4.18 for the NARX approach). These results are confirmed using the tests introduced in Section 4.4.1.9 in Tables J.106 and J.107 for the MISO DMDc approach, tables J.108 and J.109 for the MISO NARX approach, tables J.110 and J.111 for the MIMO DMDc approach, and tables J.112 and J.113 for the MIMO NARX approach.

Using the first attempt results presented in Section 4.4.1.9 as reference, it is shown that even if there are still distribution residuals with heavy-tailed, non-Gaussian shapes with autocorrelation components, it can be noticed that there were several improvements for both MISO and MIMO approaches. In both cases, the autocorrelation components in the DMDc model was reduced considerably (slightly better for the MISO case).

It can be concluded that the selected regressors explained in a better way the system to be modelled with respect of the first approach with no regressor selection analysis. Additionally, even if the heteroscedasticity tests failed to reject the alternative hypothesis, it was explored graphically if the variance obtained were increased in the time series representation, and for all cases the voltages residuals remained bounded at constant variance.

In the case of the obtained responses of NARX models, they also showed no relevant improvement in the revision of these assumptions. For the case of autocorrelation, one of the lags seems to be increased slightly in comparison with the first regression. Therefore, the ANN structure used is not improving the non-linear behaviour obtained by this input simplification and a more complex structure is required (i.e., increasing model order or the number of neurons/layers implied in the construction of model).

After this analysis, it can be concluded that the selected regressors are not able to fully explain the system to be modelled, and therefore, other regressor should be explored to improve the performance. Nevertheless, this is required if the model is desired to develop long-term voltage predictions.

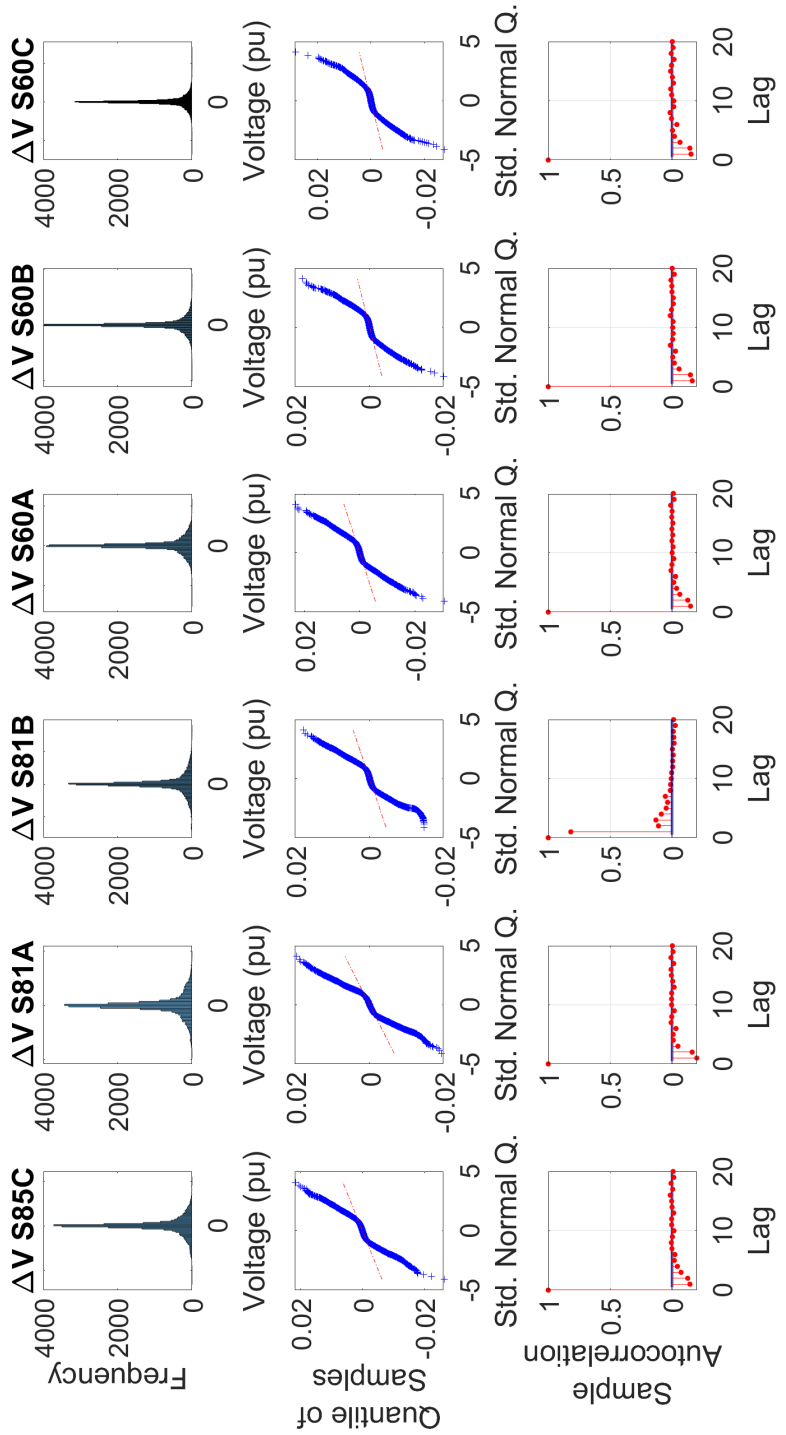


Figure 4.31: Histogram, Q-Q plot and ACF of residuals from voltages predictions using DMDc technique and training dataset of selected regressors in MISO structure

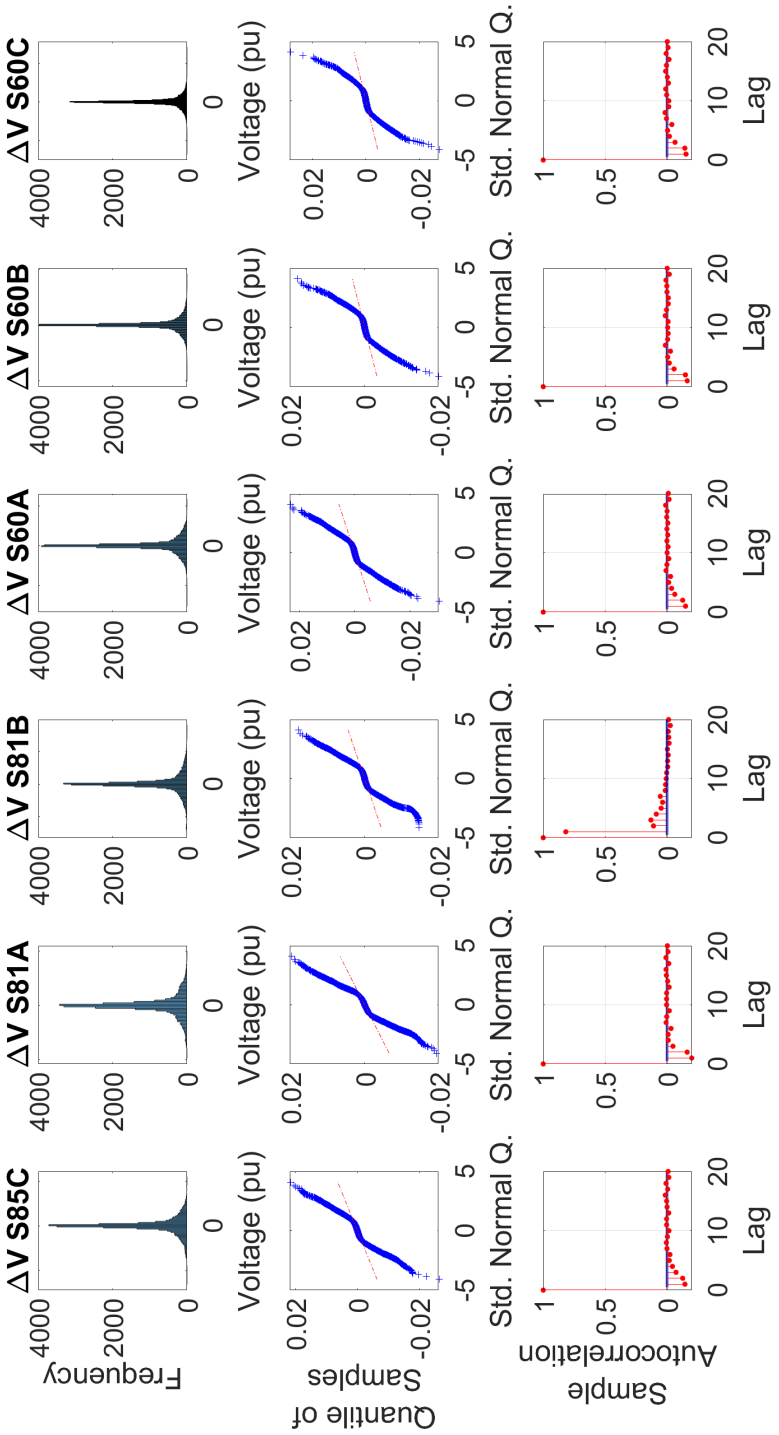


Figure 4.32: Histogram, Q-Q plot and ACF of residuals from voltages predictions using DMDc technique and validation dataset of selected regressors in MISO structure

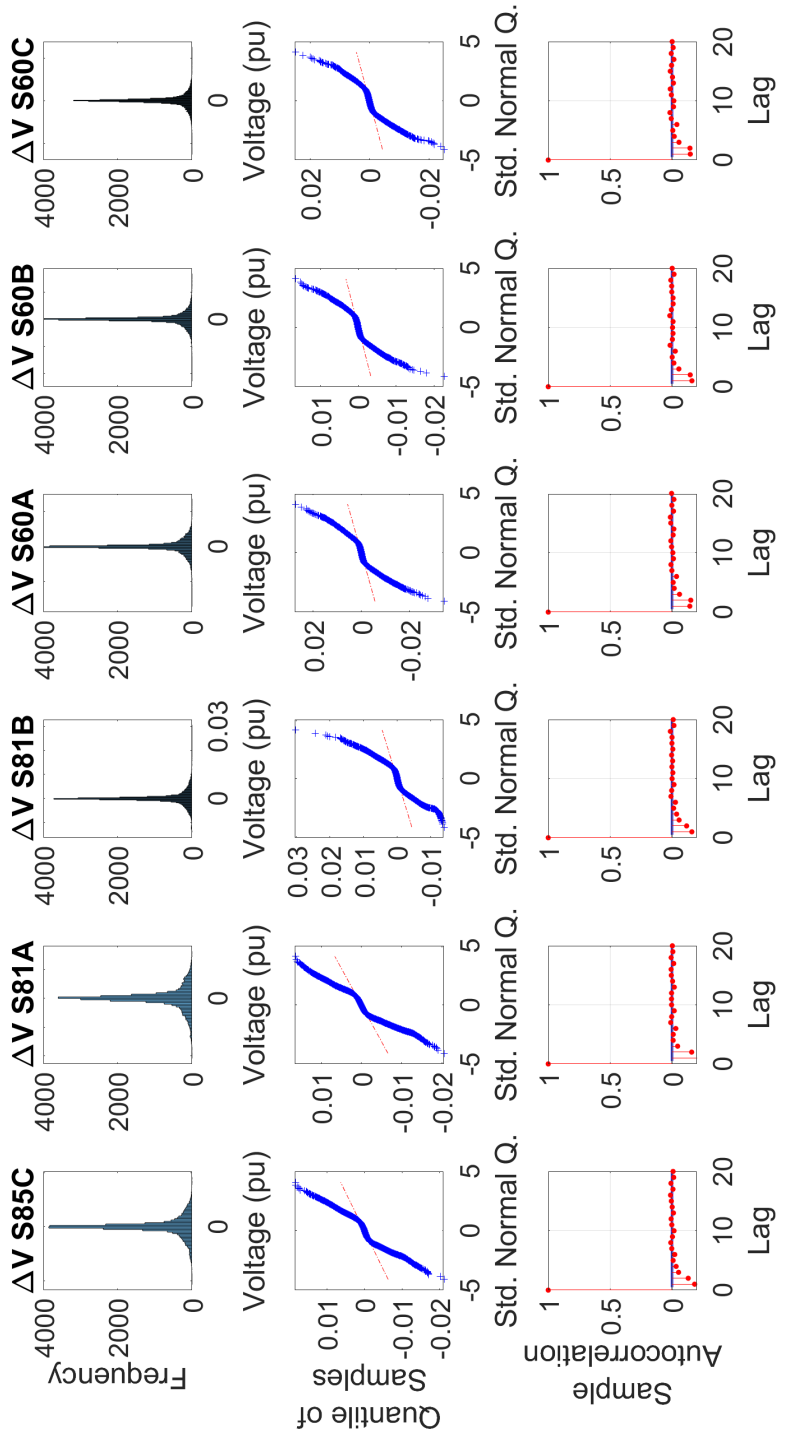


Figure 4.33: Histogram, Q-Q plot and ACF of residuals from voltages predictions using NARX technique and training dataset of selected regressors in MISO structure

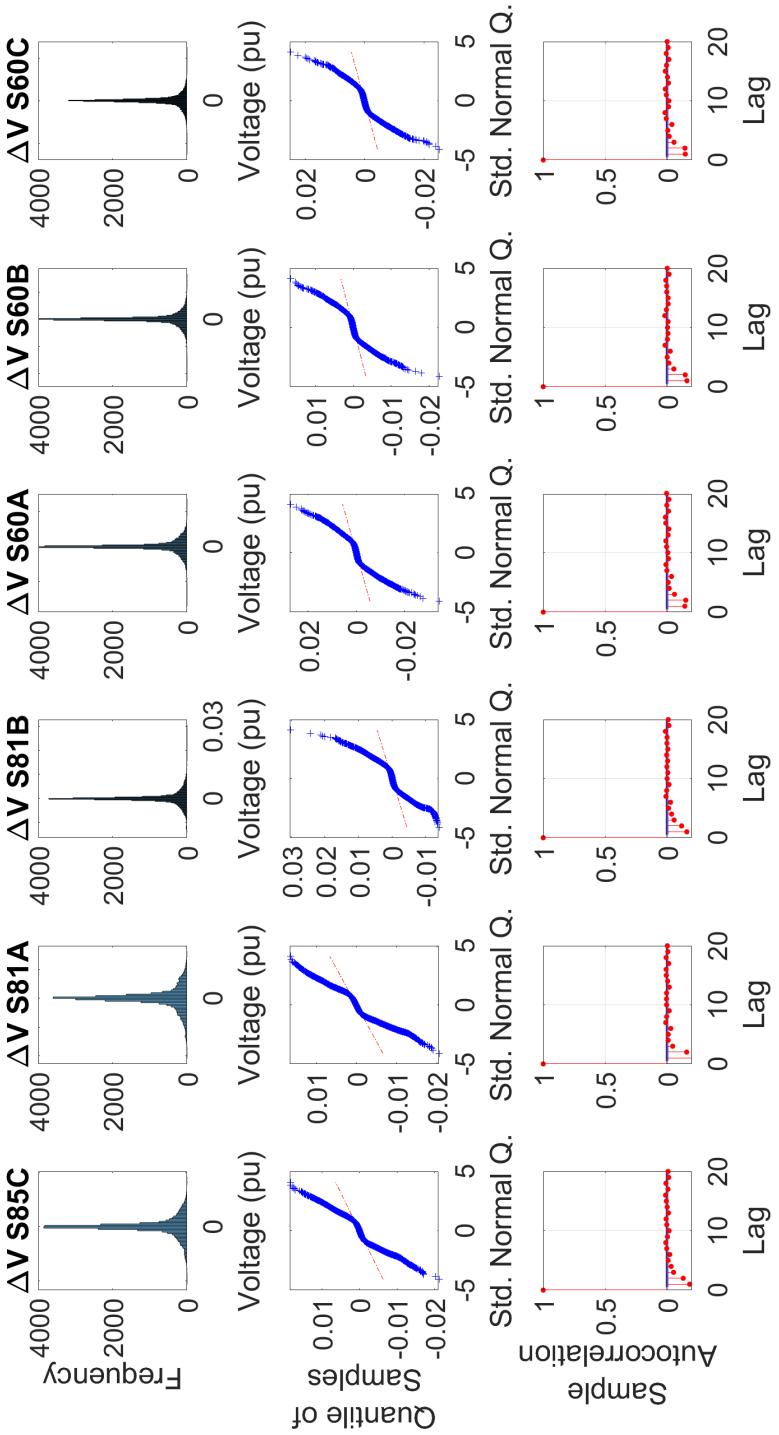


Figure 4.34: Histogram, Q-Q plot and ACF of residuals from voltages predictions using NARX technique and validation dataset of selected regressors in MISO structure

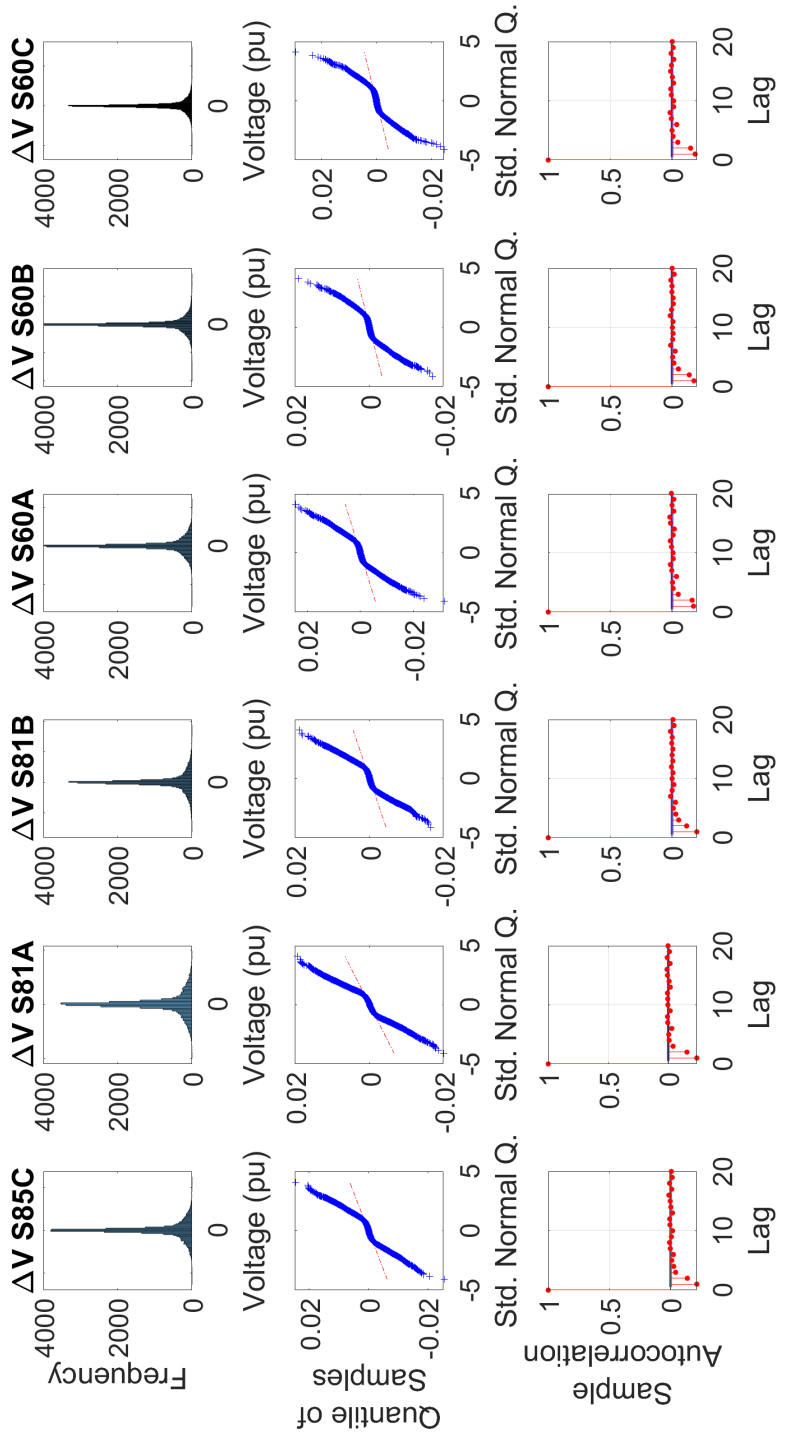


Figure 4.35: Histogram, Q-Q plot and ACF of residuals from voltages predictions using DMDc technique and training dataset of selected regressors in MIMO structure

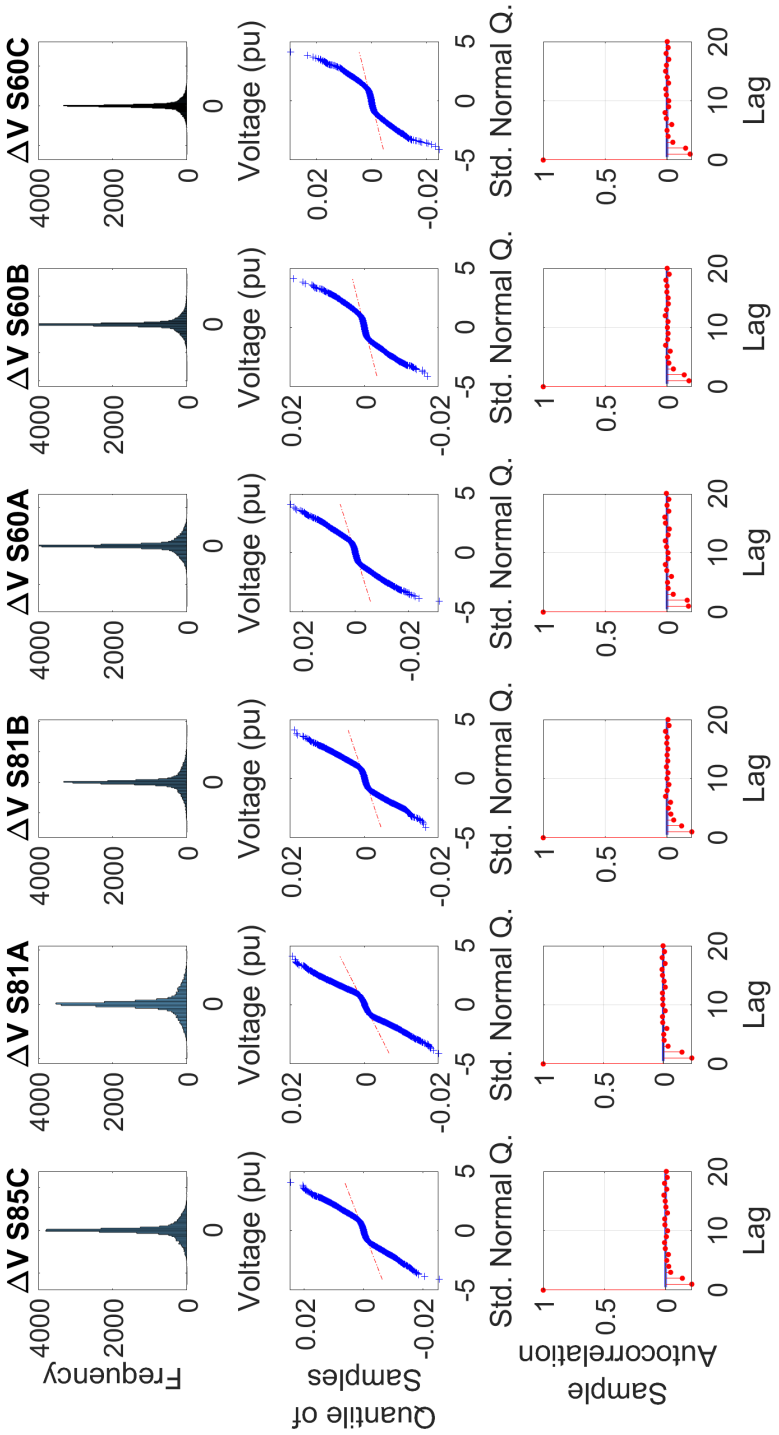


Figure 4.36: Histogram, Q-Q plot and ACF of residuals from voltages predictions using DMDc technique and validation dataset of selected regressors in MIMO structure

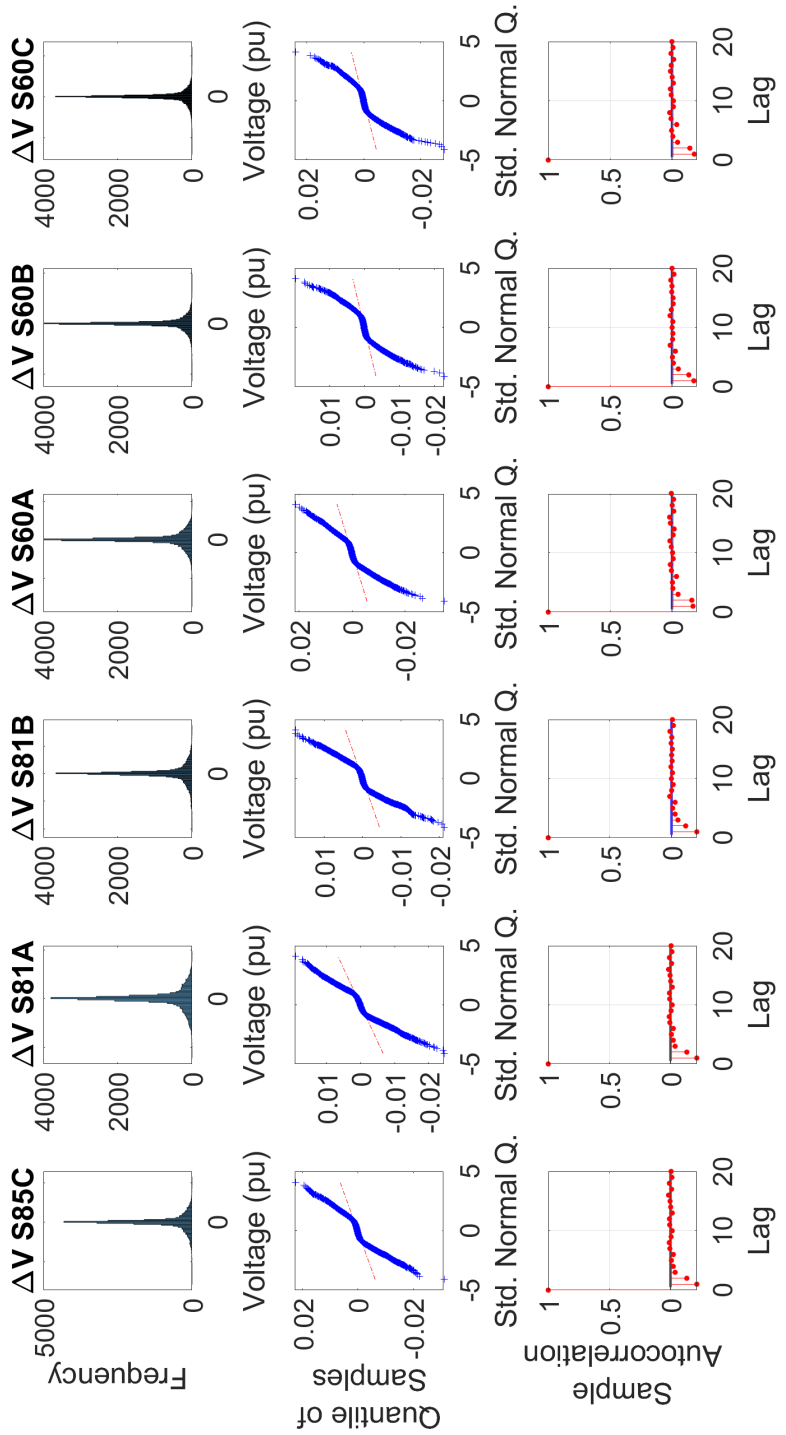


Figure 4.37: Histogram, Q-Q plot and ACF of residuals from voltages predictions using NARX technique and training dataset of selected regressors in MIMO structure

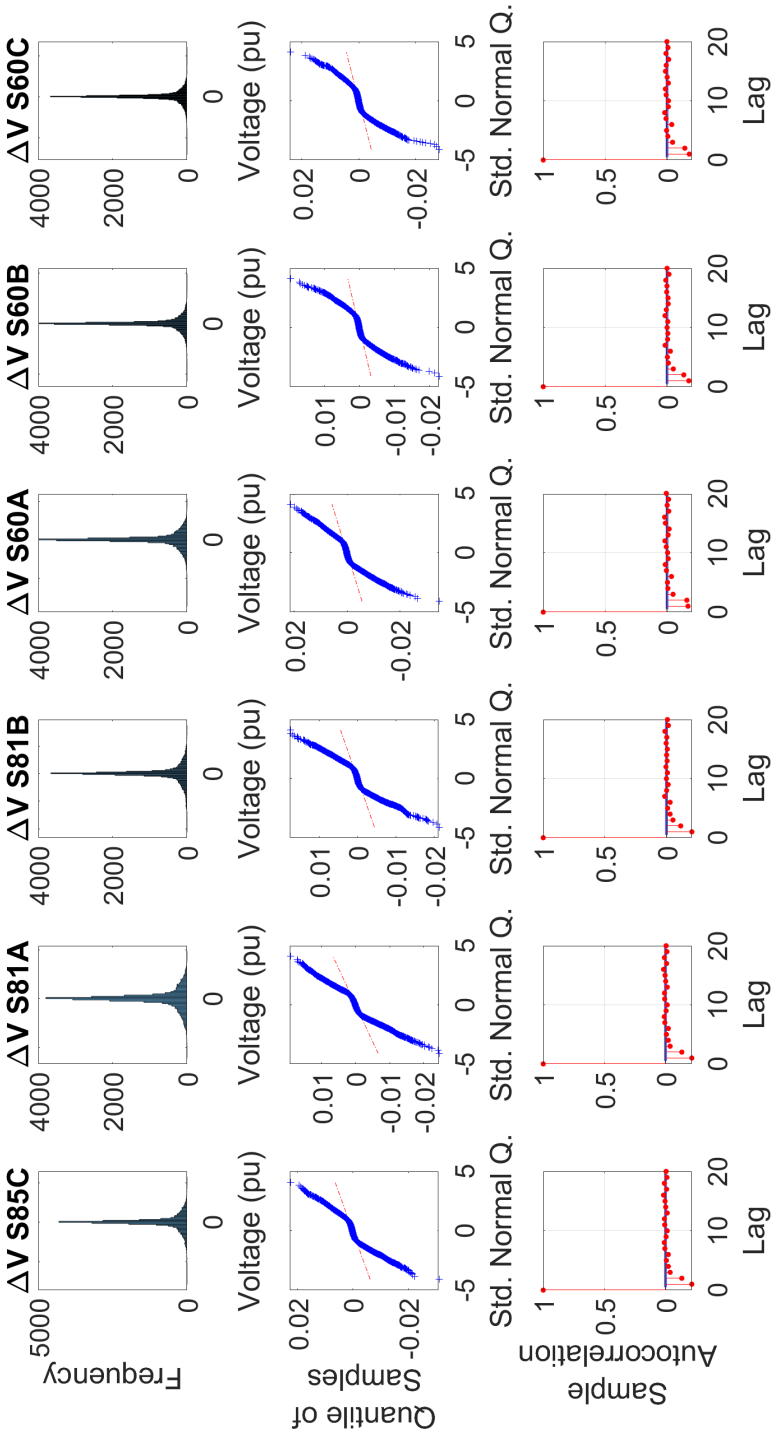


Figure 4.38: Histogram, Q-Q plot and ACF of residuals from voltages predictions using NARX technique and validation dataset of selected regressors in MIMO structure

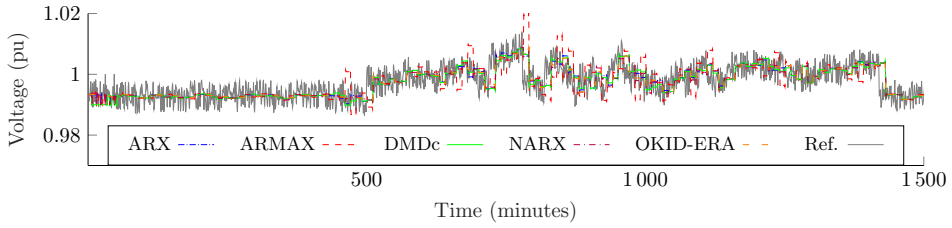
For this purpose, it is shown that it is still doing a good work for 1-step ahead, and the obtained models are stable with all the advantages of a reduced-order linear representation. This means that even if the results were not fully as expected, they are able to represent the voltage dynamics to predict using the proposed metrics, which are based purely in measurements.

To complement the obtained results, models and predictions are normally followed by a prediction interval supported statistically speaking to offer some compensation for the inaccuracies obtained in models. This interval might results useful, especially when models are applied in the context of robust or stochastic control, where those intervals can be used to inform control decisions. The definition of these intervals and the procedure is explained in section 4.7.

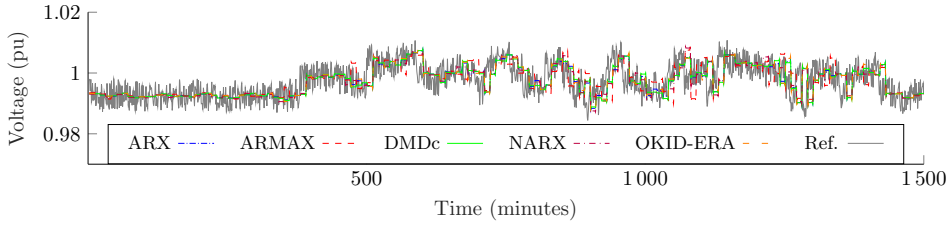
4.5 Validation of obtained models

In order to evaluate the performance of the obtained models using reference data, a comparison of the responses is done with the original data set of measurements obtained every 1-minute. A portion of the results obtained from the MISO structure and the predicted voltages are presented in Figures 4.39 and 4.40 from the original training and validation datasets with 1-minute resolution, respectively. Table 4.25 shows the performance for training on each node; Table 4.26 shows the performance for validation on each node. In a similar way, a portion of the results obtained from the MIMO structure and the predicted voltages are presented in Figures 4.41 and 4.42 from the original training and validation dataset with 1-minute resolution, respectively. Table 4.27 and 4.28 show the performance for both datasets.

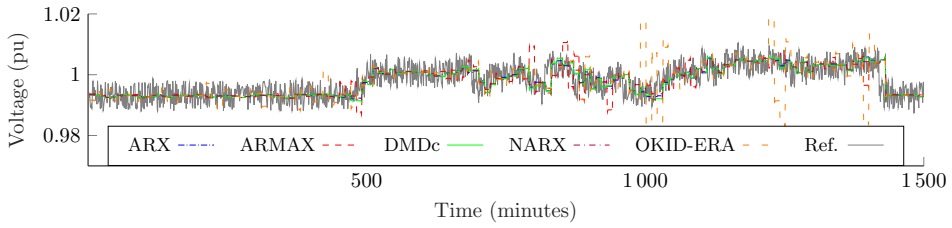
Since the model provides responses with a 10-minute resolution, it is assumed that the obtained voltage will remain constant every 10 minutes, while the reference data is changing every minute. It is shown that the performance metrics are reduced compared to the same dataset using a 10-minute resolution, as there are voltage oscillations in the reference signal with higher resolution. Nevertheless, both graphical responses for the obtained results shown in tables confirm that both model structures are consistent with the obtained responses, providing good performance for all the models (with the exception of OKID-ERA, which still showed difficulties in the tuning process). This shows that the models were able to follow the nature of voltage oscillation with higher resolution, even if the data is measured every 10 minutes.



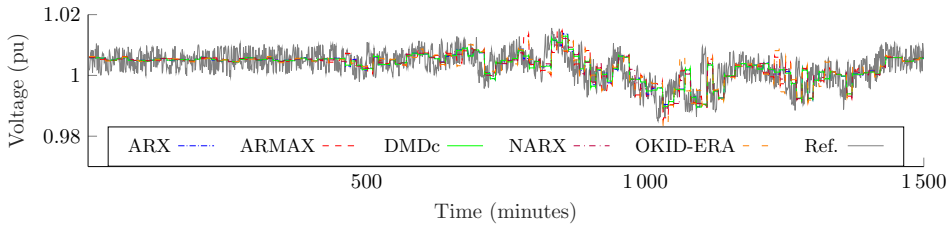
(a) Node S85C



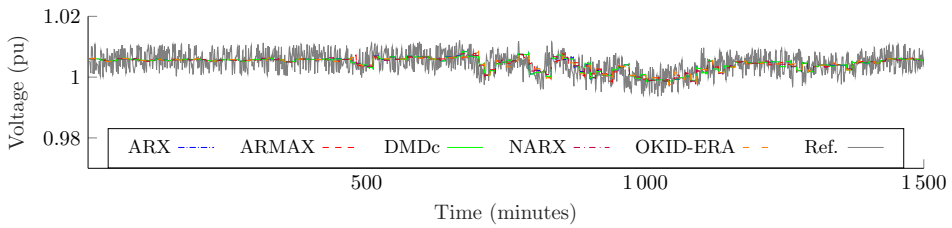
(b) Node S81A



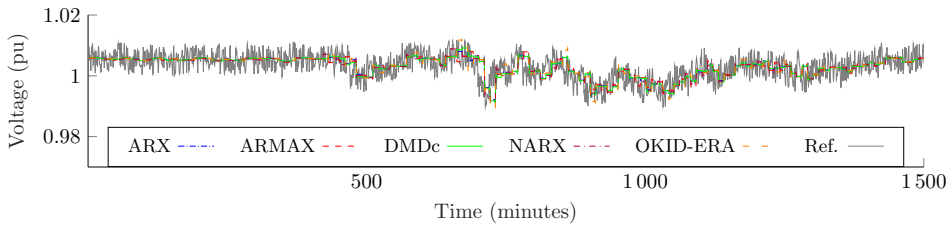
(c) Node S81B



(d) Node S60A

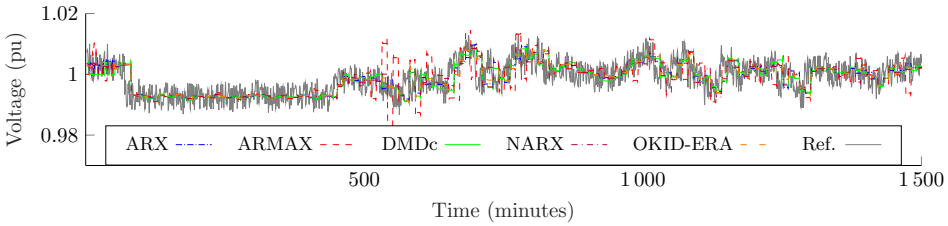


(e) Node S60B

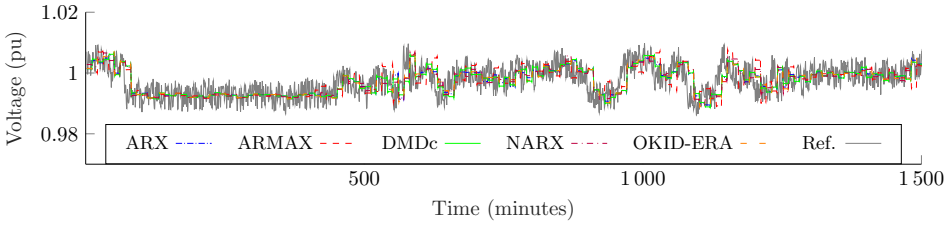


(f) Node S60C

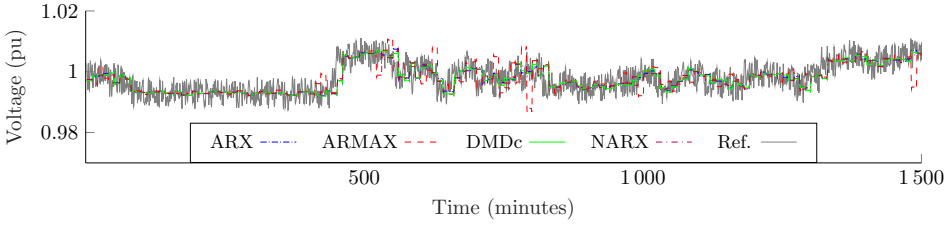
Figure 4.39: Portion of voltage predictions 1 step ahead using selected regressors training dataset in MISO structure (1-minute resolution)



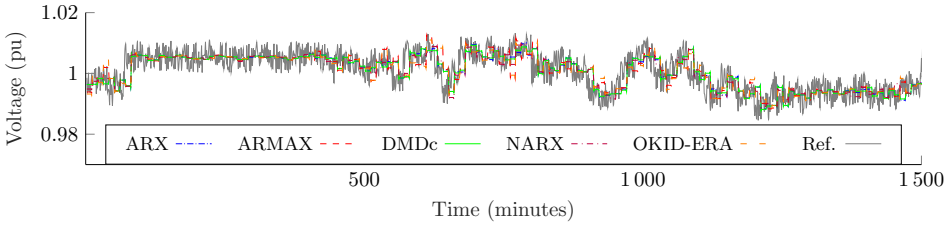
(a) Node S85C



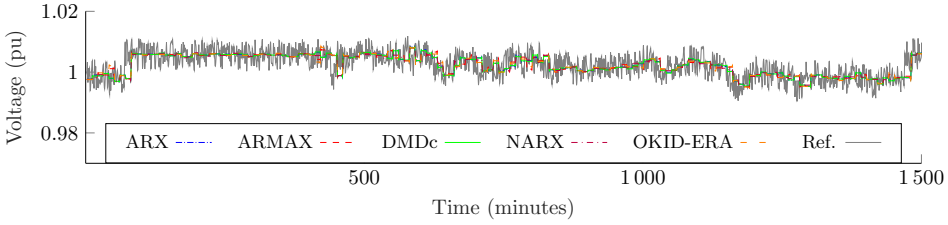
(b) Node S81A



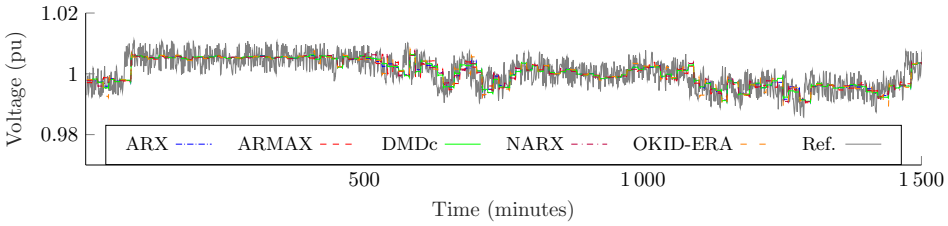
(c) Node S81B



(d) Node S60A



(e) Node S60B



(f) Node S60C

Figure 4.40: Portion of voltage predictions 1 step ahead using selected regressors validation dataset in MISO structure (1-minute resolution)

Table 4.25: Results of models for voltage prediction on each measured node using selected regressors training dataset in MISO structure (1-minute resolution)

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.45	0.16	0.45	0.47	0.45
R^2 S81A	0.28	0.14	0.28	0.27	0.26
R^2 S81B	0.44	0.18	0.44	-2.67	0.44
R^2 S60A	0.40	0.35	0.40	0.32	0.39
R^2 S60B	0.40	0.38	0.40	0.37	0.39
R^2 S60C	0.43	0.43	0.43	0.37	0.42
NRMSE S85C	0.12	0.15	0.12	0.12	0.12
NRMSE S81A	0.14	0.15	0.14	0.14	0.14
NRMSE S82B	0.13	0.19	0.13	0.49	0.13
NRMSE S60A	0.09	0.10	0.09	0.10	0.09
NRMSE S60B	0.08	0.08	0.08	0.08	0.08
NRMSE S60C	0.08	0.08	0.08	0.08	0.08
AIC S85C	-3.43e6	-3.30e6	-3.43e6	-3.44e6	-3.43e6
BIC S85C	-3.43e6	-3.30e6	-3.43e6	-3.44e6	-3.43e6
AIC S81A	-3.43e6	-3.38e6	-3.43e6	-3.43e6	-3.43e6
BIC S81A	-3.43e6	-3.38e6	-3.43e6	-3.43e6	-3.43e6
AIC S81B	-3.50e6	-3.38e6	-3.50e6	-2.91e6	-3.49e6
BIC S81B	-3.50e6	-3.38e6	-3.50e6	-2.91e6	-3.49e6
AIC S60A	-3.44e6	-3.41e6	-3.44e6	-3.40e6	-3.43e6
BIC S60A	-3.44e6	-3.41e6	-3.44e6	-3.40e6	-3.43e6
AIC S60B	-3.57e6	-3.56e6	-3.57e6	-3.55e6	-3.56e6
BIC S60B	-3.57e6	-3.56e6	-3.57e6	-3.55e6	-3.56e6
AIC S60C	-3.52e6	-3.52e6	-3.52e6	-3.49e6	-3.52e6
BIC S60C	-3.52e6	-3.52e6	-3.52e6	-3.49e6	-3.52e6

Table 4.27: Results of models for voltage prediction using selected regressors training dataset in MIMO structure (1-minute resolution)

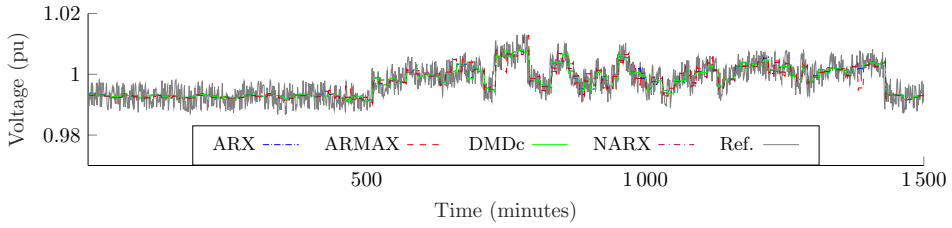
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.45	0.40	0.45	-4.29	0.45
R^2 S81A	0.25	0.19	0.26	-14.99	0.25
R^2 S81B	0.43	0.36	0.43	-1.07	0.43
R^2 S60A	0.40	0.33	0.40	-18.54	0.40
R^2 S60B	0.39	0.35	0.39	-3.80	0.39
R^2 S60C	0.43	0.41	0.43	-6.61	0.43
NRMSE S85C	0.12	0.12	0.12	0.36	0.12
NRMSE S81A	0.14	0.15	0.14	0.66	0.14
NRMSE S82B	0.13	0.14	0.13	0.26	0.13
NRMSE S60A	0.09	0.10	0.09	0.52	0.09
NRMSE S60B	0.08	0.08	0.08	0.22	0.08
NRMSE S60C	0.08	0.08	0.08	0.29	0.08
AIC	-3.48e6	-3.45e6	-3.48e6	-2.71e6	-3.48e6
BIC	-3.48e6	-3.45e6	-3.48e6	-2.71e6	-3.48e6

Table 4.26: Results of models for voltage prediction on each measured node using selected regressors validation dataset in MISO structure (1-minute resolution)

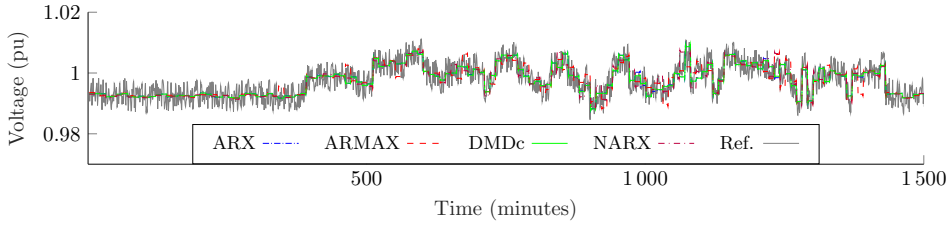
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.45	0.14	0.45	0.46	0.44
R^2 S81A	0.27	0.13	0.27	0.26	0.26
R^2 S81B	0.43	0.15	0.42	-2.85	0.42
R^2 S60A	0.38	0.33	0.39	0.30	0.38
R^2 S60B	0.39	0.37	0.39	0.36	0.38
R^2 S60C	0.42	0.42	0.42	0.36	0.41
NRMSE S85C	0.12	0.14	0.12	0.11	0.12
NRMSE S81A	0.14	0.15	0.14	0.14	0.14
NRMSE S82B	0.14	0.17	0.14	0.35	0.14
NRMSE S60A	0.10	0.11	0.10	0.11	0.10
NRMSE S60B	0.08	0.09	0.08	0.09	0.09
NRMSE S60C	0.10	0.10	0.10	0.10	0.10
AIC S85C	-3.43e6	-3.29e6	-3.43e6	-3.44e6	-3.43e6
BIC S85C	-3.43e6	-3.29e6	-3.43e6	-3.44e6	-3.43e6
AIC S81A	-3.43e6	-3.38e6	-3.43e6	-3.43e6	-3.42e6
BIC S81A	-3.43e6	-3.38e6	-3.43e6	-3.43e6	-3.42e6
AIC S81B	-3.49e6	-3.37e6	-3.49e6	-2.90e6	-3.49e6
BIC S81B	-3.49e6	-3.37e6	-3.49e6	-2.90e6	-3.49e6
AIC S60A	-3.43e6	-3.41e6	-3.44e6	-3.40e6	-3.43e6
BIC S60A	-3.43e6	-3.41e6	-3.44e6	-3.40e6	-3.43e6
AIC S60B	-3.56e6	-3.55e6	-3.56e6	-3.55e6	-3.56e6
BIC S60B	-3.56e6	-3.55e6	-3.56e6	-3.55e6	-3.56e6
AIC S60C	-3.52e6	-3.52e6	-3.52e6	-3.49e6	-3.51e6
BIC S60C	-3.52e6	-3.52e6	-3.52e6	-3.49e6	-3.51e6

Table 4.28: Results of models for voltage prediction using selected regressors validation dataset in MIMO structure (1-minute resolution)

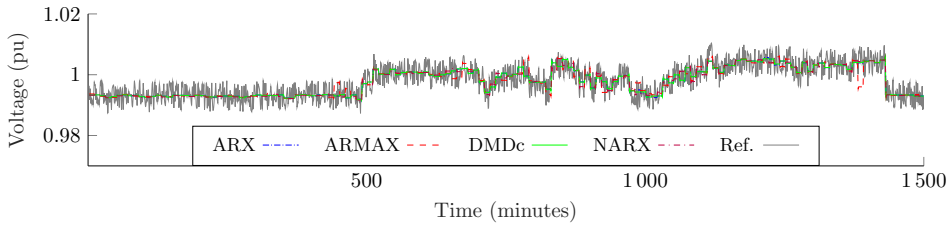
Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.44	0.39	0.44	-4.65	0.44
R^2 S81A	0.25	0.19	0.25	-16.34	0.24
R^2 S81B	0.41	0.35	0.41	-1.20	0.41
R^2 S60A	0.38	0.31	0.38	-20.19	0.38
R^2 S60B	0.38	0.34	0.38	-4.11	0.38
R^2 S60C	0.42	0.40	0.42	-7.14	0.42
NRMSE S85C	0.12	0.12	0.12	0.38	0.12
NRMSE S81A	0.14	0.15	0.14	0.69	0.14
NRMSE S82B	0.14	0.14	0.14	0.26	0.14
NRMSE S60A	0.10	0.11	0.10	0.61	0.10
NRMSE S60B	0.08	0.09	0.08	0.24	0.08
NRMSE S60C	0.10	0.10	0.10	0.37	0.10
AIC	-3.47e6	-3.45e6	-3.47e6	-2.69e6	-3.47e6
BIC	-3.47e6	-3.45e6	-3.47e6	-2.69e6	-3.47e6



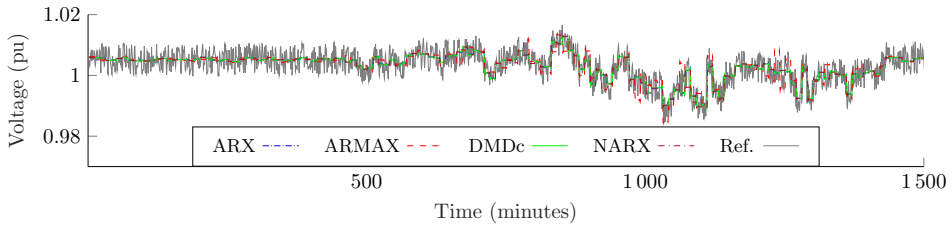
(a) Node S85C



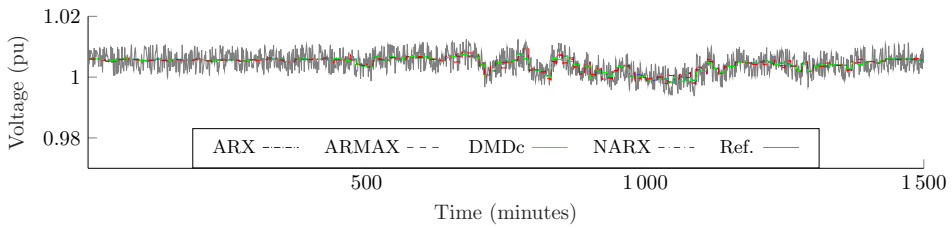
(b) Node S81A



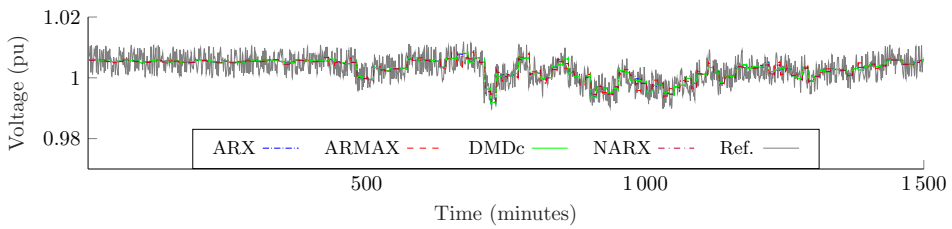
(c) Node S81B



(d) Node S60A

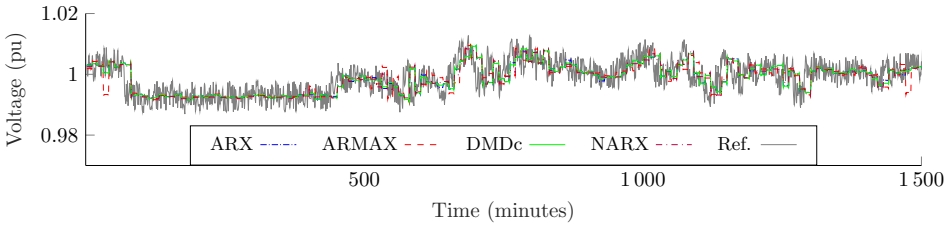


(e) Node S60B

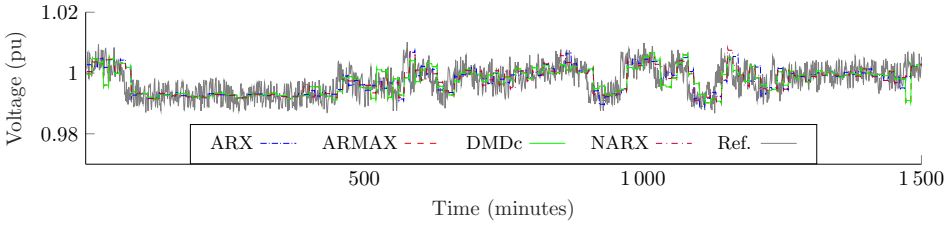


(f) Node S60C

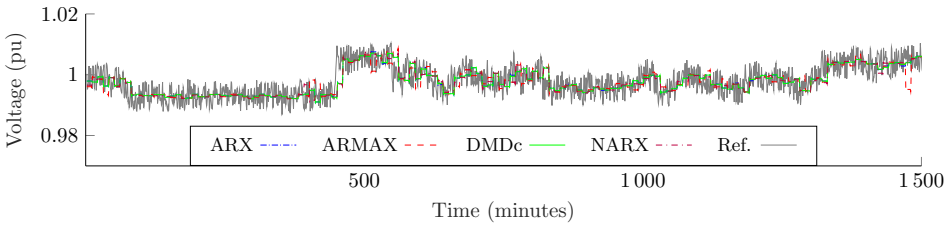
Figure 4.41: Portion of voltage predictions 1 step ahead using selected regressors training dataset in MIMO structure (1-minute resolution)



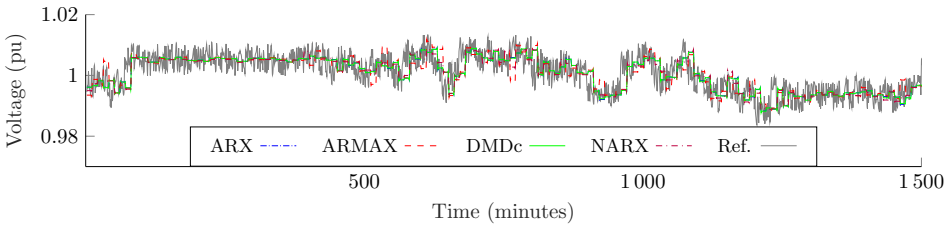
(a) Node S85C



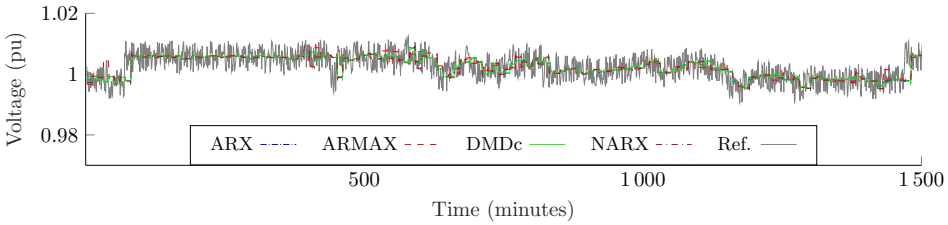
(b) Node S81A



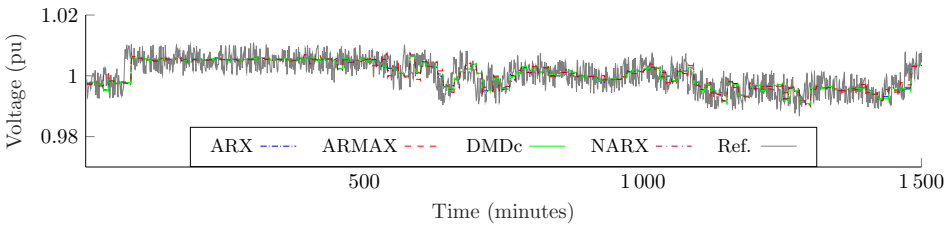
(c) Node S81B



(d) Node S60A



(e) Node S60B



(f) Node S60C

Figure 4.42: Portion of voltage predictions 1 step ahead using selected regressors validation dataset in MIMO structure (1-minute resolution)

4.6 Obtained results for "n-step ahead" predictions

The previous results show that the initial assumptions of the results were not met, and it is not possible to draw conclusive findings about the performance of the models for different step-ahead horizons. Nevertheless, an evaluation of responses was conducted for different horizons to assess the capability of the obtained model to predict the response several steps into the future (beyond 1 step ahead).

It is shown that even though there is no way to guarantee the response statistically beyond one step ahead, due to the nature of the problem, the performance of the model 2 steps ahead is relatively good. For step 3, the performance drops considerably in some of the nodes. For longer horizons (more than 6 steps or one hour), the model is not able to predict voltage variations in a reasonable way. This indicates that this approach is not meant to be used for long-horizon predictions, but rather for short-term predictions.

Table 4.29: Results of models for 2-steps voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.45	0.27	0.44	0.42	0.43
R^2 S81A	0.18	0.10	0.17	0.03	0.17
R^2 S81B	0.47	0.28	0.47	-5.11	0.47
R^2 S60A	0.37	0.33	0.36	0.28	0.35
R^2 S60B	0.52	0.51	0.52	0.49	0.52
R^2 S60C	0.51	0.50	0.49	0.50	0.48
NRMSE S85C	0.14	0.16	0.14	0.14	0.14
NRMSE S81A	0.18	0.19	0.18	0.19	0.18
NRMSE S82B	0.16	0.19	0.16	0.55	0.16
NRMSE S60A	0.10	0.10	0.10	0.11	0.10
NRMSE S60B	0.07	0.07	0.07	0.07	0.07
NRMSE S60C	0.07	0.07	0.07	0.07	0.08
AIC S85C	-3.51e5	-3.43e5	-3.51e5	-3.50e5	-3.50e5
BIC S85C	-3.51e5	-3.43e5	-3.51e5	-3.50e5	-3.50e5
AIC S81A	-3.51e5	-3.48e5	-3.51e5	-3.46e5	-3.51e5
BIC S81A	-3.51e5	-3.48e5	-3.51e5	-3.46e5	-3.51e5
AIC S81B	-3.63e5	-3.53e5	-3.62e5	-2.86e5	-3.62e5
BIC S81B	-3.63e5	-3.53e5	-3.62e5	-2.86e5	-3.62e5
AIC S60A	-3.52e5	-3.50e5	-3.52e5	-3.48e5	-3.51e5
BIC S60A	-3.52e5	-3.50e5	-3.52e5	-3.48e5	-3.51e5
AIC S60B	-3.80e5	-3.79e5	-3.80e5	-3.78e5	-3.80e5
BIC S60B	-3.80e5	-3.79e5	-3.80e5	-3.78e5	-3.80e5
AIC S60C	-3.69e5	-3.69e5	-3.68e5	-3.69e5	-3.68e5
BIC S60C	-3.69e5	-3.69e5	-3.68e5	-3.69e5	-3.68e5

Table 4.30: Results of models for 2-steps voltage prediction on each measured node using selected regressors training dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.44	0.41	0.45	-5.11	0.42
R^2 S81A	0.19	0.13	0.19	-14.72	0.14
R^2 S81B	0.47	0.41	0.47	-2.97	0.45
R^2 S60A	0.37	0.33	0.37	-26.85	0.35
R^2 S60B	0.52	0.50	0.52	-6.50	0.51
R^2 S60C	0.50	0.49	0.50	-11.93	0.49
NRMSE S85C	0.14	0.14	0.14	0.46	0.14
NRMSE S81A	0.18	0.18	0.18	0.78	0.18
NRMSE S82B	0.16	0.17	0.16	0.44	0.16
NRMSE S60A	0.10	0.10	0.10	0.68	0.10
NRMSE S60B	0.07	0.07	0.07	0.27	0.07
NRMSE S60C	0.07	0.07	0.07	0.38	0.07
AIC	-3.61e5	-3.59e5	-3.61e5	-2.72e5	-3.60e5
BIC	-3.61e5	-3.59e5	-3.61e5	-2.72e5	-3.60e5

Table 4.31: Results of models for 2-steps voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.43	0.25	0.43	0.41	0.42
R^2 S81A	0.16	0.09	0.15	0.01	0.15
R^2 S81B	0.45	0.24	0.44	-5.39	0.44
R^2 S60A	0.35	0.31	0.35	0.27	0.34
R^2 S60B	0.49	0.48	0.50	0.46	0.49
R^2 S60C	0.49	0.49	0.47	0.49	0.47
NRMSE S85C	0.14	0.16	0.14	0.14	0.14
NRMSE S81A	0.17	0.18	0.17	0.19	0.17
NRMSE S82B	0.17	0.19	0.17	0.56	0.17
NRMSE S60A	0.12	0.12	0.12	0.13	0.12
NRMSE S60B	0.08	0.08	0.08	0.08	0.08
NRMSE S60C	0.10	0.10	0.10	0.10	0.10
AIC S85C	-3.51e5	-3.42e5	-3.50e5	-3.49e5	-3.50e5
BIC S85C	-3.51e5	-3.42e5	-3.50e5	-3.49e5	-3.50e5
AIC S81A	-3.51e5	-3.48e5	-3.50e5	-3.46e5	-3.50e5
BIC S81A	-3.51e5	-3.48e5	-3.50e5	-3.46e5	-3.50e5
AIC S81B	-3.61e5	-3.52e5	-3.61e5	-2.85e5	-3.61e5
BIC S81B	-3.61e5	-3.51e5	-3.61e5	-2.85e5	-3.61e5
AIC S60A	-3.52e5	-3.50e5	-3.52e5	-3.48e5	-3.51e5
BIC S60A	-3.52e5	-3.50e5	-3.52e5	-3.48e5	-3.51e5
AIC S60B	-3.78e5	-3.78e5	-3.78e5	-3.76e5	-3.78e5
BIC S60B	-3.78e5	-3.78e5	-3.78e5	-3.76e5	-3.78e5
AIC S60C	-3.69e5	-3.69e5	-3.68e5	-3.69e5	-3.67e5
BIC S60C	-3.69e5	-3.69e5	-3.68e5	-3.69e5	-3.67e5

Table 4.32: Results of models for 2-steps voltage prediction on each measured node using selected regressors validation dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.43	0.40	0.43	-5.55	0.41
R^2 S81A	0.17	0.12	0.18	-16.11	0.13
R^2 S81B	0.44	0.38	0.44	-3.29	0.44
R^2 S60A	0.35	0.31	0.35	-30.20	0.34
R^2 S60B	0.50	0.48	0.50	-7.09	0.50
R^2 S60C	0.49	0.48	0.49	-13.24	0.48
NRMSE S85C	0.14	0.14	0.14	0.47	0.15
NRMSE S81A	0.17	0.18	0.17	0.78	0.19
NRMSE S82B	0.17	0.18	0.17	0.46	0.17
NRMSE S60A	0.12	0.12	0.12	0.82	0.11
NRMSE S60B	0.08	0.08	0.08	0.31	0.08
NRMSE S60C	0.10	0.10	0.10	0.52	0.08
AIC	-3.60e5	-3.59e5	-3.60e5	-2.69e5	-3.59e5
BIC	-3.60e5	-3.59e5	-3.60e5	-2.69e5	-3.59e5

Table 4.33: Results of models for 3-steps voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.33	0.01	0.31	0.30	0.32
R^2 S81A	0.03	-0.02	0.06	-0.35	0.06
R^2 S81B	0.34	-0.12	0.35	-15.62	0.35
R^2 S60A	0.25	0.21	0.24	0.18	0.25
R^2 S60B	0.42	0.40	0.43	0.39	0.42
R^2 S60C	0.41	0.41	0.41	0.25	0.40
NRMSE S85C	0.15	0.19	0.15	0.16	0.15
NRMSE S81A	0.19	0.20	0.19	0.23	0.19
NRMSE S82B	0.18	0.24	0.18	0.91	0.18
NRMSE S60A	0.11	0.11	0.11	0.12	0.11
NRMSE S60B	0.08	0.08	0.08	0.08	0.08
NRMSE S60C	0.08	0.08	0.08	0.09	0.08
AIC S85C	-3.45e5	-3.33e5	-3.45e5	-3.44e5	-3.45e5
BIC S85C	-3.45e5	-3.33e5	-3.45e5	-3.44e5	-3.45e5
AIC S81A	-3.46e5	-3.44e5	-3.47e5	-3.35e5	-3.47e5
BIC S81A	-3.46e5	-3.44e5	-3.47e5	-3.35e5	-3.47e5
AIC S81B	-3.56e5	-3.39e5	-3.56e5	-2.54e5	-3.56e5
BIC S81B	-3.56e5	-3.39e5	-3.56e5	-2.55e5	-3.56e5
AIC S60A	-3.46e5	-3.45e5	-3.46e5	-3.43e5	-3.46e5
BIC S60A	-3.46e5	-3.45e5	-3.46e5	-3.43e5	-3.46e5
AIC S60B	-3.74e5	-3.73e5	-3.75e5	-3.72e5	-3.74e5
BIC S60B	-3.74e5	-3.73e5	-3.75e5	-3.72e5	-3.74e5
AIC S60C	-3.64e5	-3.64e5	-3.64e5	-3.56e5	-3.63e5
BIC S60C	-3.64e5	-3.64e5	-3.64e5	-3.56e5	-3.63e5

Table 4.34: Results of models for 3-steps voltage prediction on each measured node using selected regressors training dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.32	0.30	0.33	-1.74	0.30
R^2 S81A	0.04	-0.01	0.04	-3.72	0.00
R^2 S81B	0.34	0.29	0.34	-1.35	0.32
R^2 S60A	0.25	0.22	0.25	-5.05	0.23
R^2 S60B	0.42	0.40	0.42	-1.65	0.41
R^2 S60C	0.41	0.40	0.41	-2.36	0.39
NRMSE S85C	0.15	0.16	0.15	0.31	0.16
NRMSE S81A	0.19	0.20	0.19	0.43	0.20
NRMSE S82B	0.18	0.19	0.18	0.34	0.18
NRMSE S60A	0.11	0.11	0.11	0.32	0.11
NRMSE S60B	0.08	0.08	0.08	0.16	0.08
NRMSE S60C	0.08	0.08	0.08	0.19	0.08
AIC	-3.55e5	-3.54e5	-3.55e5	-3.05e5	-3.54e5
BIC	-3.55e5	-3.54e5	-3.55e5	-3.05e5	-3.54e5

Table 4.35: Results of models for 3-steps voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.31	-0.02	0.30	0.29	0.30
R^2 S81A	0.02	-0.03	0.05	-0.37	0.04
R^2 S81B	0.32	-0.17	0.32	-16.22	0.32
R^2 S60A	0.23	0.19	0.22	0.16	0.23
R^2 S60B	0.40	0.38	0.41	0.36	0.40
R^2 S60C	0.39	0.39	0.39	0.22	0.38
NRMSE S85C	0.15	0.18	0.15	0.15	0.15
NRMSE S81A	0.19	0.19	0.19	0.22	0.19
NRMSE S82B	0.18	0.24	0.18	0.93	0.18
NRMSE S60A	0.13	0.13	0.13	0.14	0.13
NRMSE S60B	0.08	0.08	0.08	0.09	0.08
NRMSE S60C	0.11	0.11	0.11	0.12	0.11
AIC S85C	-3.45e5	-3.32e5	-3.44e5	-3.44e5	-3.44e5
BIC S85C	-3.45e5	-3.32e5	-3.44e5	-3.44e5	-3.44e5
AIC S81A	-3.46e5	-3.44e5	-3.47e5	-3.35e5	-3.46e5
BIC S81A	-3.46e5	-3.44e5	-3.47e5	-3.35e5	-3.46e5
AIC S81B	-3.55e5	-3.38e5	-3.55e5	-2.54e5	-3.55e5
BIC S81B	-3.55e5	-3.38e5	-3.55e5	-2.54e5	-3.55e5
AIC S60A	-3.47e5	-3.45e5	-3.46e5	-3.44e5	-3.47e5
BIC S60A	-3.47e5	-3.45e5	-3.46e5	-3.44e5	-3.47e5
AIC S60B	-3.73e5	-3.72e5	-3.73e5	-3.71e5	-3.73e5
BIC S60B	-3.73e5	-3.72e5	-3.73e5	-3.71e5	-3.73e5
AIC S60C	-3.63e5	-3.63e5	-3.63e5	-3.56e5	-3.63e5
BIC S60C	-3.63e5	-3.63e5	-3.63e5	-3.56e5	-3.63e5

Table 4.36: Results of models for 3-steps voltage prediction on each measured node using selected regressors validation dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.32	0.30	0.33	-1.74	0.30
R^2 S81A	0.04	-0.01	0.04	-3.72	0.00
R^2 S81B	0.34	0.29	0.34	-1.35	0.32
R^2 S60A	0.25	0.22	0.25	-5.05	0.23
R^2 S60B	0.42	0.40	0.42	-1.65	0.41
R^2 S60C	0.41	0.40	0.41	-2.36	0.39
NRMSE S85C	0.15	0.16	0.15	0.31	0.16
NRMSE S81A	0.19	0.20	0.19	0.43	0.20
NRMSE S82B	0.18	0.19	0.18	0.34	0.18
NRMSE S60A	0.11	0.11	0.11	0.32	0.11
NRMSE S60B	0.08	0.08	0.08	0.16	0.08
NRMSE S60C	0.08	0.08	0.08	0.19	0.08
AIC	-3.55e5	-3.54e5	-3.55e5	-3.05e5	-3.54e5
BIC	-3.55e5	-3.54e5	-3.55e5	-3.05e5	-3.54e5

Table 4.37: Results of models for 6-steps voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.08	-0.21	0.08	0.10	-0.02
R^2 S81A	-0.24	-0.23	-0.22	-0.39	-0.37
R^2 S81B	0.07	0.05	0.10	-13.25	0.00
R^2 S60A	0.05	-0.06	0.02	0.05	-0.11
R^2 S60B	0.22	0.18	0.23	0.24	0.16
R^2 S60C	0.21	0.22	0.22	0.19	0.07
NRMSE S85C	0.18	0.20	0.18	0.18	0.19
NRMSE S81A	0.22	0.22	0.22	0.23	0.23
NRMSE S82B	0.21	0.22	0.21	0.84	0.22
NRMSE S60A	0.12	0.13	0.13	0.12	0.14
NRMSE S60B	0.09	0.09	0.09	0.09	0.09
NRMSE S60C	0.09	0.09	0.09	0.09	0.10
AIC S85C	-3.36e5	-3.27e5	-3.35e5	-3.36e5	-3.32e5
BIC S85C	-3.36e5	-3.27e5	-3.35e5	-3.36e5	-3.32e5
AIC S81A	-3.38e5	-3.38e5	-3.39e5	-3.35e5	-3.35e5
BIC S81A	-3.38e5	-3.38e5	-3.39e5	-3.34e5	-3.35e5
AIC S81B	-3.45e5	-3.44e5	-3.46e5	-2.60e5	-3.43e5
BIC S81B	-3.45e5	-3.44e5	-3.46e5	-2.59e5	-3.43e5
AIC S60A	-3.39e5	-3.35e5	-3.38e5	-3.39e5	-3.34e5
BIC S60A	-3.39e5	-3.35e5	-3.38e5	-3.39e5	-3.34e5
AIC S60B	-3.65e5	-3.63e5	-3.65e5	-3.66e5	-3.62e5
BIC S60B	-3.65e5	-3.63e5	-3.65e5	-3.66e5	-3.62e5
AIC S60C	-3.55e5	-3.55e5	-3.55e5	-3.54e5	-3.49e5
BIC S60C	-3.55e5	-3.55e5	-3.55e5	-3.54e5	-3.49e5

Table 4.38: Results of models for 6-steps voltage prediction on each measured node using selected regressors training dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.09	0.10	0.09	-2.01	-0.04
R^2 S81A	-0.23	-0.22	-0.23	-7.54	-0.45
R^2 S81B	0.06	0.08	0.06	-2.52	-0.05
R^2 S60A	0.03	0.02	0.03	-9.53	-0.11
R^2 S60B	0.21	0.21	0.21	-1.97	0.11
R^2 S60C	0.22	0.21	0.22	-3.20	0.10
NRMSE S85C	0.18	0.18	0.18	0.32	0.19
NRMSE S81A	0.22	0.22	0.22	0.57	0.24
NRMSE S82B	0.22	0.21	0.22	0.42	0.23
NRMSE S60A	0.13	0.13	0.13	0.42	0.14
NRMSE S60B	0.09	0.09	0.09	0.17	0.09
NRMSE S60C	0.09	0.09	0.09	0.21	0.10
AIC	-3.46e5	-3.46e5	-3.46e5	-2.95e5	-3.42e5
BIC	-3.46e5	-3.46e5	-3.46e5	-2.95e5	-3.42e5

Table 4.39: Results of models for 6-steps voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.06	-0.23	0.05	0.08	-0.04
R^2 S81A	-0.27	-0.26	-0.24	-0.42	-0.40
R^2 S81B	0.03	0.00	0.06	-13.83	-0.04
R^2 S60A	0.03	-0.09	0.00	0.03	-0.14
R^2 S60B	0.19	0.14	0.20	0.21	0.12
R^2 S60C	0.19	0.20	0.20	0.16	0.04
NRMSE S85C	0.18	0.20	0.18	0.18	0.19
NRMSE S81A	0.21	0.21	0.21	0.23	0.22
NRMSE S82B	0.22	0.22	0.22	0.86	0.23
NRMSE S60A	0.15	0.15	0.15	0.15	0.16
NRMSE S60B	0.10	0.10	0.10	0.10	0.10
NRMSE S60C	0.12	0.12	0.12	0.13	0.14
AIC S85C	-3.35e5	-3.26e5	-3.35e5	-3.35e5	-3.32e5
BIC S85C	-3.35e5	-3.26e5	-3.35e5	-3.35e5	-3.32e5
AIC S81A	-3.38e5	-3.38e5	-3.38e5	-3.34e5	-3.35e5
BIC S81A	-3.38e5	-3.38e5	-3.38e5	-3.34e5	-3.35e5
AIC S81B	-3.44e5	-3.43e5	-3.45e5	-2.59e5	-3.42e5
BIC S81B	-3.44e5	-3.43e5	-3.45e5	-2.59e5	-3.42e5
AIC S60A	-3.39e5	-3.36e5	-3.38e5	-3.39e5	-3.34e5
BIC S60A	-3.39e5	-3.36e5	-3.38e5	-3.39e5	-3.34e5
AIC S60B	-3.64e5	-3.62e5	-3.64e5	-3.64e5	-3.61e5
BIC S60B	-3.64e5	-3.62e5	-3.64e5	-3.64e5	-3.61e5
AIC S60C	-3.54e5	-3.55e5	-3.55e5	-3.53e5	-3.49e5
BIC S60C	-3.54e5	-3.55e5	-3.55e5	-3.53e5	-3.49e5

Table 4.40: Results of models for 6-steps voltage prediction on each measured node using selected regressors validation dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	0.07	0.08	0.07	-2.07	-0.05
R^2 S81A	-0.26	-0.25	-0.26	-7.84	-0.46
R^2 S81B	0.02	0.04	0.02	-2.78	-0.06
R^2 S60A	0.00	0.00	0.00	-10.56	-0.12
R^2 S60B	0.18	0.17	0.18	-2.09	0.10
R^2 S60C	0.19	0.19	0.20	-3.54	0.09
NRMSE S85C	0.18	0.18	0.18	0.32	0.20
NRMSE S81A	0.21	0.21	0.21	0.56	0.25
NRMSE S82B	0.22	0.22	0.22	0.43	0.24
NRMSE S60A	0.15	0.15	0.15	0.50	0.15
NRMSE S60B	0.10	0.10	0.10	0.19	0.10
NRMSE S60C	0.12	0.12	0.12	0.29	0.11
AIC	-3.45e5	-3.46e5	-3.45e5	-2.93e5	-3.42e5
BIC	-3.45e5	-3.46e5	-3.45e5	-2.93e5	-3.42e5

Table 4.41: Results of models for 12-steps voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.21	-0.47	-0.17	-0.21	-0.25
R^2 S81A	-0.53	-0.53	-0.51	-0.90	-0.59
R^2 S81B	-0.29	-0.36	-0.26	-33.24	-0.33
R^2 S60A	-0.23	-0.35	-0.27	-0.26	-0.32
R^2 S60B	-0.05	-0.10	-0.05	-0.06	-0.10
R^2 S60C	-0.06	-0.05	-0.04	-0.24	-0.10
NRMSE S85C	0.21	0.23	0.20	0.21	0.21
NRMSE S81A	0.24	0.24	0.24	0.27	0.25
NRMSE S82B	0.25	0.26	0.25	1.30	0.26
NRMSE S60A	0.14	0.15	0.14	0.14	0.15
NRMSE S60B	0.10	0.10	0.10	0.10	0.10
NRMSE S60C	0.11	0.11	0.11	0.12	0.11
AIC S85C	-3.27e5	-3.21e5	-3.28e5	-3.27e5	-3.26e5
BIC S85C	-3.27e5	-3.21e5	-3.28e5	-3.27e5	-3.26e5
AIC S81A	-3.31e5	-3.31e5	-3.32e5	-3.25e5	-3.30e5
BIC S81A	-3.31e5	-3.31e5	-3.32e5	-3.25e5	-3.30e5
AIC S81B	-3.35e5	-3.33e5	-3.35e5	-2.32e5	-3.34e5
BIC S81B	-3.35e5	-3.33e5	-3.35e5	-2.32e5	-3.34e5
AIC S60A	-3.31e5	-3.28e5	-3.30e5	-3.30e5	-3.29e5
BIC S60A	-3.31e5	-3.28e5	-3.30e5	-3.30e5	-3.29e5
AIC S60B	-3.55e5	-3.54e5	-3.55e5	-3.55e5	-3.54e5
BIC S60B	-3.55e5	-3.54e5	-3.55e5	-3.55e5	-3.54e5
AIC S60C	-3.46e5	-3.46e5	-3.46e5	-3.40e5	-3.44e5
BIC S60C	-3.45e5	-3.46e5	-3.46e5	-3.40e5	-3.44e5

Table 4.42: Results of models for 12-steps voltage prediction on each measured node using selected regressors training dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.22	-0.22	-0.22	-0.84	-0.24
R^2 S81A	-0.53	-0.52	-0.52	-2.24	-0.55
R^2 S81B	-0.30	-0.30	-0.30	-1.30	-0.31
R^2 S60A	-0.26	-0.27	-0.26	-1.53	-0.30
R^2 S60B	-0.07	-0.07	-0.07	-0.88	-0.10
R^2 S60C	-0.05	-0.05	-0.05	-0.34	-0.08
NRMSE S85C	0.21	0.21	0.21	0.25	0.21
NRMSE S81A	0.24	0.24	0.24	0.35	0.24
NRMSE S82B	0.25	0.25	0.25	0.34	0.25
NRMSE S60A	0.14	0.14	0.14	0.20	0.15
NRMSE S60B	0.10	0.10	0.10	0.14	0.11
NRMSE S60C	0.11	0.11	0.11	0.12	0.11
AIC	-3.37e5	-3.37e5	-3.37e5	-3.20e5	-3.36e5
BIC	-3.37e5	-3.37e5	-3.37e5	-3.20e5	-3.36e5

Table 4.43: Results of models for 12-steps voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.22	-0.49	-0.18	-0.22	-0.26
R^2 S81A	-0.56	-0.57	-0.54	-0.95	-0.63
R^2 S81B	-0.33	-0.41	-0.30	-35.29	-0.37
R^2 S60A	-0.26	-0.38	-0.29	-0.28	-0.35
R^2 S60B	-0.08	-0.12	-0.07	-0.08	-0.12
R^2 S60C	-0.07	-0.06	-0.06	-0.26	-0.12
NRMSE S85C	0.20	0.22	0.20	0.20	0.21
NRMSE S81A	0.24	0.24	0.24	0.26	0.24
NRMSE S82B	0.26	0.26	0.25	1.35	0.26
NRMSE S60A	0.17	0.17	0.17	0.17	0.17
NRMSE S60B	0.11	0.11	0.11	0.11	0.11
NRMSE S60C	0.14	0.14	0.14	0.15	0.15
AIC S85C	-3.27e5	-3.20e5	-3.28e5	-3.27e5	-3.26e5
BIC S85C	-3.27e5	-3.20e5	-3.28e5	-3.26e5	-3.26e5
AIC S81A	-3.31e5	-3.31e5	-3.32e5	-3.24e5	-3.30e5
BIC S81A	-3.31e5	-3.31e5	-3.32e5	-3.24e5	-3.30e5
AIC S81B	-3.34e5	-3.32e5	-3.35e5	-2.31e5	-3.33e5
BIC S81B	-3.34e5	-3.32e5	-3.35e5	-2.31e5	-3.33e5
AIC S60A	-3.31e5	-3.28e5	-3.30e5	-3.30e5	-3.29e5
BIC S60A	-3.31e5	-3.28e5	-3.30e5	-3.30e5	-3.29e5
AIC S60B	-3.55e5	-3.53e5	-3.55e5	-3.54e5	-3.53e5
BIC S60B	-3.54e5	-3.53e5	-3.55e5	-3.54e5	-3.53e5
AIC S60C	-3.46e5	-3.46e5	-3.46e5	-3.40e5	-3.44e5
BIC S60C	-3.46e5	-3.46e5	-3.46e5	-3.40e5	-3.44e5

Table 4.44: Results of models for 12-steps voltage prediction on each measured node using selected regressors validation dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.23	-0.23	-0.23	-0.87	-0.25
R^2 S81A	-0.56	-0.55	-0.56	-2.32	-0.56
R^2 S81B	-0.34	-0.33	-0.34	-1.41	-0.32
R^2 S60A	-0.29	-0.29	-0.29	-1.69	-0.31
R^2 S60B	-0.09	-0.10	-0.09	-0.91	-0.11
R^2 S60C	-0.07	-0.07	-0.06	-0.36	-0.09
NRMSE S85C	0.20	0.20	0.20	0.25	0.22
NRMSE S81A	0.24	0.24	0.24	0.35	0.25
NRMSE S82B	0.26	0.26	0.26	0.35	0.26
NRMSE S60A	0.17	0.17	0.17	0.24	0.16
NRMSE S60B	0.11	0.11	0.11	0.15	0.12
NRMSE S60C	0.14	0.14	0.14	0.16	0.12
AIC	-3.37e5	-3.37e5	-3.37e5	-3.20e5	-3.36e5
BIC	-3.37e5	-3.37e5	-3.37e5	-3.20e5	-3.36e5

Table 4.45: Results of models for 144-steps voltage prediction on each measured node using selected regressors training dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.04	-0.39	-0.04	-0.05	-0.08
R^2 S81A	-0.32	-0.32	-0.30	-0.72	-0.36
R^2 S81B	-0.13	-0.37	-0.10	-172.76	-0.16
R^2 S60A	-0.08	-0.19	-0.11	-0.12	-0.15
R^2 S60B	-0.10	-0.14	-0.09	-0.10	-0.14
R^2 S60C	0.02	0.03	0.04	0.00	-0.01
NRMSE S85C	0.19	0.22	0.19	0.19	0.19
NRMSE S81A	0.23	0.23	0.22	0.26	0.23
NRMSE S82B	0.24	0.26	0.23	2.93	0.24
NRMSE S60A	0.13	0.14	0.14	0.14	0.14
NRMSE S60B	0.11	0.11	0.10	0.11	0.11
NRMSE S60C	0.10	0.10	0.10	0.10	0.10
AIC S85C	-3.31e5	-3.22e5	-3.31e5	-3.31e5	-3.30e5
BIC S85C	-3.31e5	-3.22e5	-3.31e5	-3.31e5	-3.30e5
AIC S81A	-3.36e5	-3.35e5	-3.36e5	-3.27e5	-3.34e5
BIC S81A	-3.35e5	-3.35e5	-3.36e5	-3.27e5	-3.34e5
AIC S81B	-3.38e5	-3.32e5	-3.39e5	-1.81e5	-3.37e5
BIC S81B	-3.38e5	-3.32e5	-3.39e5	-1.81e5	-3.37e5
AIC S60A	-3.34e5	-3.31e5	-3.33e5	-3.33e5	-3.32e5
BIC S60A	-3.34e5	-3.31e5	-3.33e5	-3.33e5	-3.32e5
AIC S60B	-3.53e5	-3.52e5	-3.53e5	-3.53e5	-3.52e5
BIC S60B	-3.53e5	-3.52e5	-3.53e5	-3.53e5	-3.52e5
AIC S60C	-3.47e5	-3.48e5	-3.48e5	-3.47e5	-3.46e5
BIC S60C	-3.47e5	-3.47e5	-3.48e5	-3.47e5	-3.46e5

Table 4.46: Results of models for 144-steps voltage prediction on each measured node using selected regressors training dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.05	-0.05	-0.05	-3.50	-0.08
R^2 S81A	-0.31	-0.31	-0.31	-6.73	-0.37
R^2 S81B	-0.14	-0.14	-0.14	-1.58	-0.17
R^2 S60A	-0.11	-0.12	-0.11	-3.71	-0.13
R^2 S60B	-0.11	-0.12	-0.11	-2.66	-0.14
R^2 S60C	0.03	0.02	0.03	-1.95	0.01
NRMSE S85C	0.19	0.19	0.19	0.40	0.19
NRMSE S81A	0.22	0.22	0.22	0.55	0.23
NRMSE S82B	0.24	0.24	0.24	0.36	0.24
NRMSE S60A	0.13	0.14	0.13	0.28	0.14
NRMSE S60B	0.11	0.11	0.11	0.19	0.11
NRMSE S60C	0.10	0.10	0.10	0.18	0.10
AIC	-3.40e5	-3.40e5	-3.40e5	-2.99e5	-3.39e5
BIC	-3.40e5	-3.40e5	-3.40e5	-2.99e5	-3.39e5

Table 4.47: Results of models for 144-steps voltage prediction on each measured node using selected regressors validation dataset in MISO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.04	-0.39	-0.04	-0.05	-0.08
R^2 S81A	-0.32	-0.32	-0.30	-0.72	-0.36
R^2 S81B	-0.13	-0.37	-0.10	-172.76	-0.16
R^2 S60A	-0.08	-0.19	-0.11	-0.12	-0.15
R^2 S60B	-0.10	-0.14	-0.09	-0.10	-0.14
R^2 S60C	0.02	0.03	0.04	0.00	-0.01
NRMSE S85C	0.19	0.22	0.19	0.19	0.19
NRMSE S81A	0.23	0.23	0.22	0.26	0.23
NRMSE S82B	0.24	0.26	0.23	2.93	0.24
NRMSE S60A	0.13	0.14	0.14	0.14	0.14
NRMSE S60B	0.11	0.11	0.10	0.11	0.11
NRMSE S60C	0.10	0.10	0.10	0.10	0.10
AIC S85C	-3.31e5	-3.22e5	-3.31e5	-3.31e5	-3.30e5
BIC S85C	-3.31e5	-3.22e5	-3.31e5	-3.31e5	-3.30e5
AIC S81A	-3.36e5	-3.35e5	-3.36e5	-3.27e5	-3.34e5
BIC S81A	-3.35e5	-3.35e5	-3.36e5	-3.27e5	-3.34e5
AIC S81B	-3.38e5	-3.32e5	-3.39e5	-1.81e5	-3.37e5
BIC S81B	-3.38e5	-3.32e5	-3.39e5	-1.81e5	-3.37e5
AIC S60A	-3.34e5	-3.31e5	-3.33e5	-3.33e5	-3.32e5
BIC S60A	-3.34e5	-3.31e5	-3.33e5	-3.33e5	-3.32e5
AIC S60B	-3.53e5	-3.52e5	-3.53e5	-3.53e5	-3.52e5
BIC S60B	-3.53e5	-3.52e5	-3.53e5	-3.53e5	-3.52e5
AIC S60C	-3.47e5	-3.48e5	-3.48e5	-3.47e5	-3.46e5
BIC S60C	-3.47e5	-3.47e5	-3.48e5	-3.47e5	-3.46e5

Table 4.48: Results of models for 144-steps voltage prediction on each measured node using selected regressors validation dataset in MIMO structure

Characteristics	ARX (SS)	ARMAX (SS)	DMDc	OKID-ERA	NARX
R^2 S85C	-0.05	-0.05	-0.05	-3.64	-0.09
R^2 S81A	-0.36	-0.36	-0.36	-7.11	-0.38
R^2 S81B	-0.13	-0.13	-0.13	-1.68	-0.18
R^2 S60A	-0.15	-0.16	-0.15	-3.94	-0.14
R^2 S60B	-0.16	-0.17	-0.16	-2.77	-0.15
R^2 S60C	0.02	0.01	0.02	-2.03	0.00
NRMSE S85C	0.19	0.19	0.19	0.39	0.20
NRMSE S81A	0.22	0.22	0.22	0.54	0.24
NRMSE S82B	0.24	0.24	0.24	0.37	0.25
NRMSE S60A	0.16	0.16	0.16	0.33	0.15
NRMSE S60B	0.12	0.12	0.12	0.21	0.12
NRMSE S60C	0.14	0.14	0.14	0.24	0.11
AIC	-3.39e5	-3.39e5	-3.39e5	-2.98e5	-3.39e5
BIC	-3.39e5	-3.39e5	-3.39e5	-2.98e5	-3.39e5

4.7 Obtaining prediction interval for the time-series modelling

From previous steps, it is obtained a linear representation that represents the voltage quasi-dynamics based on the selected regressors after analysis. Last step of Algorithm 4.1 (Step 5), corresponds to obtaining prediction interval for the time-series modelling. Normally, time-series representations come with a prediction interval that gives a statistical boundary in which the obtained values lie with a specified probability [246]. According to Equations (4.4) and (4.5), the prediction interval is not for the predicted voltage, but it is constructed linear model that relates the differenced regressors and voltages.

A prediction interval gives an interval within which the predicted value $\overline{\Delta y}(t_{k+1})$ from equation (4.5) is expected to lie with a specified probability. This value is commonly given for a 95% prediction interval for the h-step horizon [246]. This margin can be calculated by using pre-defined values from standard distributions (normal, t-distribution, etc.). Since the distributions obtained for residuals are not following the expected standard distribution shape, it is part of the main challenge of this part of the thesis to use another technique that fits with empirical residuals.

For any modelling approach, the uncertainty associated with the prediction is product of the error associated with model parameters and the unpredictable errors from external causes [288]. The obtained heavy-tailed distribution in residuals would not increase the effect of the errors associated with the estimators in the lin-

ear regression [289]. This is because the error is averaged out in the least-squares estimate. Nevertheless, this will affect the prediction interval and it is required to explore the margin of the prediction interval obtained under these conditions.

One of the most common techniques to obtain this unpredictable error is applying bootstrapping in the residuals [290–293]. Bootstrapping is one of the broader resampling methods, in which any test or metric relies on random sampling with replacement, and at the same time assuming that the data that have not been selected are the test dataset. This procedure is repeated several times and compute the average score as estimation of model performance. However, the basic bootstrap depends on the initial sample consisting of independent and identically distributed random variables draws from a fixed population distribution. In practice and as was obtained before, residuals showed correlation components.

Another possibility is cross-validation resampling without replacement [238]. This procedure splits the training data into k parts (that is also called k-fold cross validation). Then, it is assumed that $k-1$ parts are used for training and the other for testing/validating. The procedure is repeated k times taking different part of the data each time. Finally, the average of the k scores is calculated as performance estimation. In time series, there are issues that must be considered [294]; the prediction horizon is affected according to the portion of dataset that is taken, and there will be issues with the data before the selected training data (because the prediction would be given to "past values" instead). Finally, the method can suffer from variance or bias according to the size of the fold.

The possible solution for this in time-series data is a mixture of both techniques under special considerations. In this work, two techniques were implemented to obtain the prediction intervals:

- Time-Series Split Cross-Validation (TSSCV) [295]: In this method, the cross-validation is done considering the historical correlation of data. A small portion of the variables is selected to validate, while the rest is taken for training. Figure 4.43 summarises the representation of the selected dataset on each iteration. The horizontal axis shows the training set size while the vertical axis shows the cross-validation iterations. The folds used for training and validation are depicted in blue and orange, respectively. The horizontal axis represents the time progression line that have not shuffled the dataset and maintained the chronological order. Therefore, time series is split into two folds at each iteration, where validation set is always ahead of the training set.

- Blocked Cross-Validation (BCV) [296]: This method considers leakages from

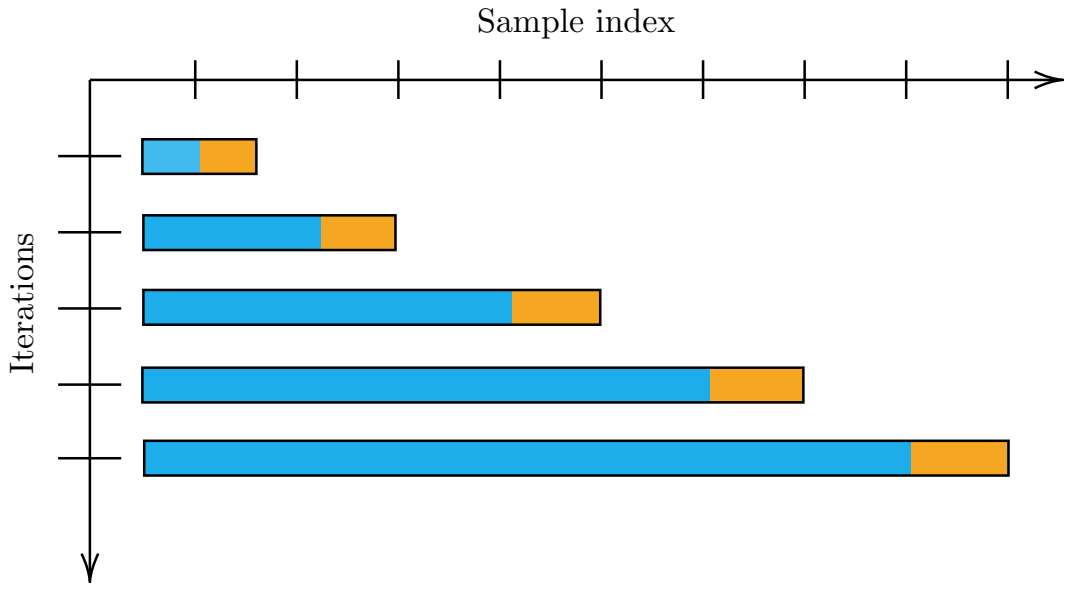


Figure 4.43: Time-series split and selection process of TSSCV method for prediction interval for training dataset (blue bar) and validation dataset (orange bar)

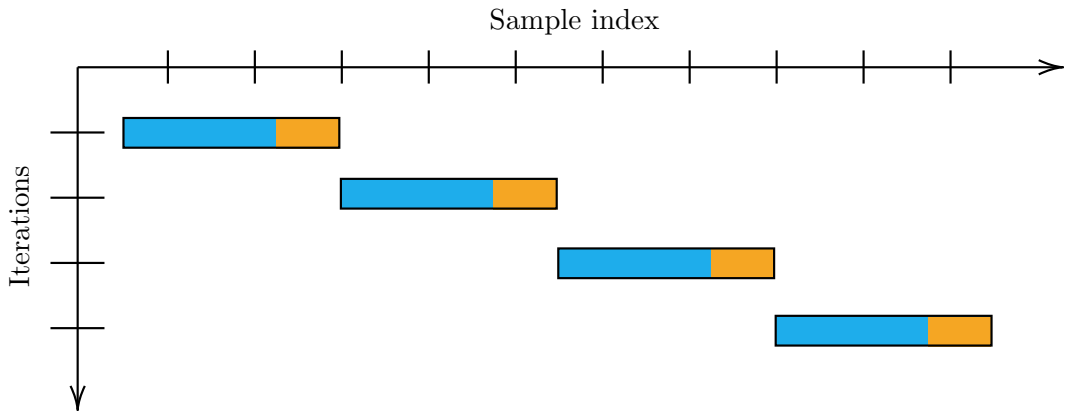


Figure 4.44: Time-series split and selection process of BCV method for prediction interval for training dataset (blue bar) and validation dataset (orange bar)

future data to the model, and therefore, future patterns are observed. Figure 4.44 summarises the selection of data, following a similar presentation of previous figure. In this case, two margins are added: the first between the training and validation folds (to avoid that the model observes lag values that are used twice, once as a regressor and as a response); and the second between the folds used at each iteration to avoid memorising patterns between each iteration.

The procedure to calculate the prediction interval using the methods mentioned before is presented in Algorithm 4.2.

Algorithm 4.2 Prediction interval calculation from data-driven time-series modelling approach

Input: Selected data, Number of repetition B, window resolution for each iteration fold
Output: Prediction interval
initialisation
Step 1 - Calculate $\Delta V(t_{k+1})$;
for $iter = 1 : B$ **do**
  Step 2 - Set the size of each fold
  Step 3 - Choose the time-series cross-validation method (TSSCV or BCV); choose a random initial point in the training set (uniform distribution)
  Step 4 - Run regressions using the obtained training/testing dataset and calculate the residual
end
Step 5 - Calculate percentiles 2.5 and 97.5 from the obtained residual distributions (95% confidence)
Step 6 - Calculate prediction intervals, add the values to $\Delta V(t_{k+1})$

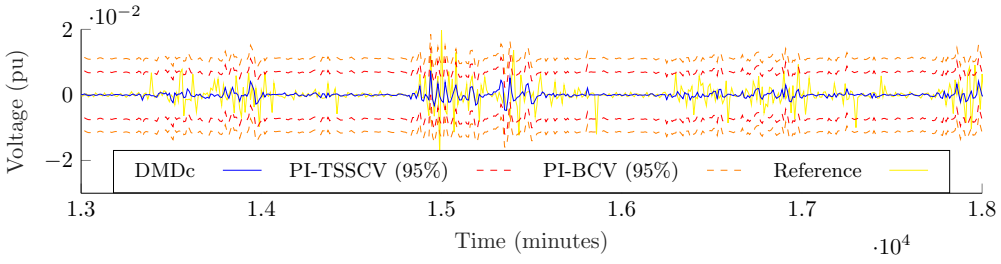
The calculation of percentiles is done using the empirical method called t-digest [297]. This technique uses a sparse representation of the ECDF of a data set, and it is useful for computing approximations of rank-based statistics (percentiles and quantiles). T-digest is used to estimate the median and any percentile from either distributed data or streaming data. The first step consists of obtaining a t-digest in each partition of the data. Once each sparse representation is "ingested", the algorithm finds from the data structure the "interesting" points of the CDF to be learned, which are called centroids, and an accumulated weight that represents the number of samples contributing to the cluster. Once the t-digest that represents the complete data set is obtained, the endpoints (or boundaries) can be estimated for each cluster and the accurate quantile is estimated by interpolating between the endpoints of each cluster.

Table 4.49: Prediction intervals and computation times obtained for the model DMDc in MISO structure

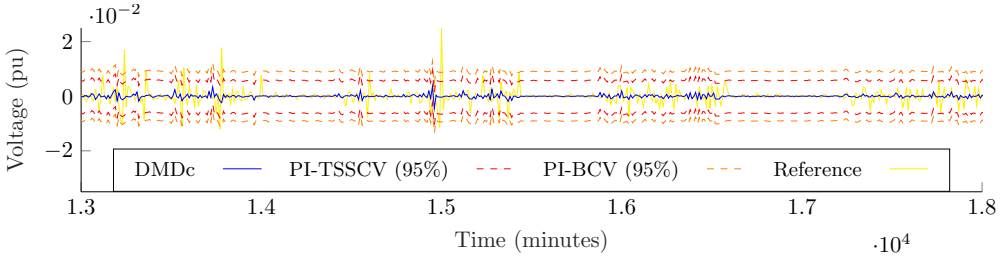
	TSSCV		BCV	
	ΔV S85C	ΔV S60C	ΔV S85C	ΔV S60C
B = 10000	[-0.0073 , 0.0070]	[-0.0063 , 0.0059]	[-0.0114 , 0.0111]	[-0.0088 , 0.0087]
$\text{Time}_{B_{10000}}$ (min)	124.77		117.69	
B = 1000	[-0.0073 , 0.0069]	[-0.0064 , 0.0059]	[-0.0113 , 0.0111]	[-0.0089 , 0.0088]
$\text{Time}_{B_{1000}}$ (min)	12.47		12.78	
B = 500	[-0.0073 , 0.0070]	[-0.002 , 0.0058]	[-0.0113 , 0.0111]	[-0.0090 , 0.0091]
$\text{Time}_{B_{500}}$ (min)	7.00		6.81	

Table 4.50: Prediction intervals and computation times obtained for the model DMDc in MIMO structure

	TSSCV		BCV	
	ΔV S85C	ΔV S60C	ΔV S85C	ΔV S60C
B = 10000	[-0.0075 , 0.0070]	[-0.0062 , 0.0058]	[-0.0105 , 0.0094]	[-0.0079 , 0.0084]
$\text{Time}_{B_{10000}}$ (min)	83.41		73.55	
B = 1000	[-0.0076 , 0.0070]	[-0.0062 , 0.0058]	[-0.0105 , 0.0095]	[-0.0078 , 0.0084]
$\text{Time}_{B_{1000}}$ (min)	8.82		7.82	



(a) ΔV S85C

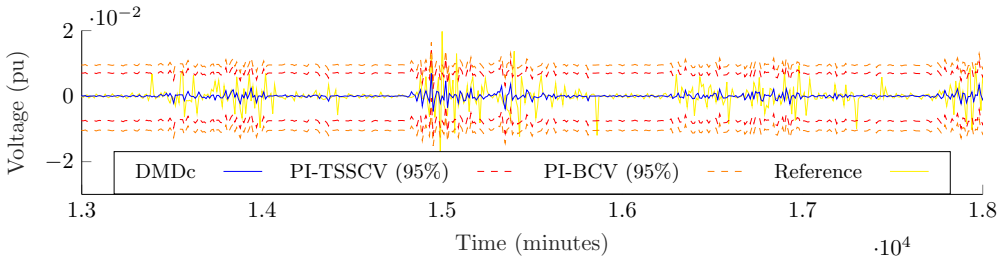


(b) ΔV S60C

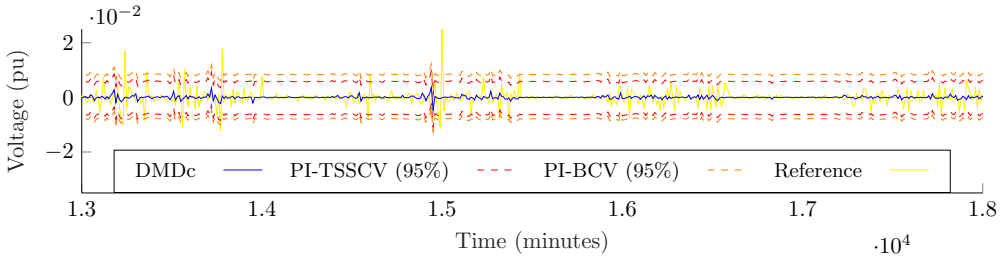
Figure 4.45: Portion of voltage variation predictions 1 step ahead and prediction intervals using selected regressors in MISO structure

To produce the residuals for both methods, the complete dataset of 1000 days were considered in as reference to be used. As illustration, results on phase C are summarised in Tables 4.49 and 4.50. A representative portion of data results are presented in Figures 4.45 and 4.46. This prediction interval is done for DMDc model. Nevertheless, same approach can be applied for each of the models.

These results show that the actual values remain within the boundaries defined in the prediction intervals for each structure. The margin obtained from BCV is wider than the one obtained from TSSCV, and both results remain within the 95% confidence level (meaning that at least 95% of cases are inside the boundaries even when there are some spikes outside of the intervals on a few occasions). The computational time for BCV is slightly lower than TSSCV. Both methods showed similar results in the obtained prediction intervals. As indicated in Tables 4.49



(a) ΔV S85C



(b) ΔV S60C

Figure 4.46: Portion of voltage variation predictions 1 step ahead and prediction intervals using selected regressors in MIMO structure

and 4.50, for the MISO case, the maximum iteration in which the response remains within the prediction time corresponds to the case of $B = 500$ repetitions, while it is about $B = 1000$ repetitions for the MIMO case. This is because the MISO case represents a model for each predicted voltage, which requires running repetitions for each one. These methods rely on the number of repetitions to guarantee the quality of the unpredicted response, but there is always a balance between the computational time (here limited in the prediction horizon) and the number of repetitions. Nevertheless, the prediction intervals for both TSSCV and BCV are similar for the lowest number of repetitions in each method to the reference cases of $B = 10000$ repetitions. In both figures, it is shown that the prediction interval is wide enough to encompass most of the predictions. Ideally, this interval should be small in order to prove that the obtained values from the model are highly accurate. However, the approach starts from the fact that there are only measurements available. If the regression is not fully able to capture all the details from the predicted values, it would result in a reduction of accuracy, which is translated into a broad prediction interval. One solution for this would be a mix of a known representation or more detailed knowledge from the distribution system that can complement the obtained information from the data-driven model. That should be translated to a reduced bandwidth in the prediction interval.

4.8 Discussion

The results obtained demonstrate that data can be effectively used to describe spatial-temporal perturbations in the distribution system. The use of M_P, M_Q, and the covariance can define relevant characteristics of the system, such as the size of perturbation and the distance and impacted nodes. By defining a covariance threshold, only closer nodes to the perturbation are considered, reducing the complexity of the model. Additionally, voltage magnitude analysis provides valuable information due to the unbalanced electromagnetic compensation in all three phases. These components can be potentially used as the input vector u and the measured voltage from selected nodes as the output vector y in the construction of control models.

The proposed methodology demonstrates that the analysis of measurable data is beneficial in constructing a good state-space model in a data-driven approach. The flexibility of this methodology allows for the use of different regression approaches and model structures. Among the linear autoregressive models, ARX models consistently showed one of the best performances in different stages and structures, while ARMAX models also produced satisfactory results. However, both methods lack a unique way to convert the model from the classical representation to the space-state representation, which can impact the dimensions of the model and computational time, especially for large and complex systems. The performance improved significantly after the reduction and simplification of selected regressors. The Koopman-operator-based method presented in this thesis (DMDc) also consistently showed similar performance to ARX, but the representation of the reduced-order model is always consistent, and the dimension of obtained models and computational time were considerably reduced. This is an advantage when a simple quick response is required to be computed. An extended revision of the size of the model is required to evaluate if the minimum size selected (number of outputs) can capture the relevant quasi-dynamics in different environments with different levels of observability. Classical subspace identification methods, such as OKID-ERA, sometimes produced better performances in some of the responses than the previously mentioned methods, but in most cases, the obtained responses did not match the reference predicted signal. The main problem with this approach is the tuning process of parameters in the algorithm to produce the model (especially when setting the "synthetic" Hankel matrix), which makes the approach difficult to be used in automation applications based on standard procedures. Additionally, computational times were not better than any of the previously obtained methods. Finally, the non-linear regression used

(NARX) showed that it can considerably fit the predicted curve in most cases and structures, but computational time is long, and it is impossible to describe anything about the internal dynamics of the obtained model. This can limit the analysis of the obtained outputs. For all the discussed models, the MISO approach showed better prediction than the MIMO case, but the trade-off is the individual analysis of each input, resulting in the production of individual models. This may increase the complexity in calculation and control approaches. Nevertheless, comparable results were obtained for both MISO and MIMO approaches. The selection of each method will depend on the chosen strategy in the control approach to be implemented.

Figures 4.45 and 4.46 also provide valuable insights into the obtained results using the proposed method. For the selected regressors, it is currently not possible to predict a big swing of power-voltage change, which is reflected in the moment when the model is not able to follow the substantial change in voltage from the reference case (which is evident as the horizon is only for one step ahead). Nevertheless, under the previous assumption of constant variance (failed in the test but observed graphically) after analyzing residuals, it is shown that even these extreme cases still fall within the prediction interval boundaries. As mentioned before, one possible solution to this issue is to explore other potential regressors that can help detect these quicker quasi-dynamics that are not currently captured in this approach.

Other possible and compatible solution would be the integration of another model that helps to explain some of the quasi-dynamics to be modelled and predicted. The proposed approach starts from the assumption that there no previous knowledge from the system that helps on predict the voltage behaviour, and therefore, only relies on available measurement. It was explored the use of different regressors to explain a possible linear representation of the variables and obtained a reduced-order model that is suitable for control. This does not make the approach incompatible with any other methods that relies on previous knowledge/models of the system. This would change slightly the structure of the model, but it will require for the algorithm to reduce the complexity of relating control and exogenous inputs and will be more focused to improve the response.

The cases presented in this thesis were critical cases, and consequently, they were the most representative to produce models. The most critical component to be aware of is the assumption in the obtained residuals. Residuals were not fulfilling initial assumptions, but according to the results, models are still good for short-term predictions, which is enough of a control horizon in a system with stochasticity in measurable regressors. The inclusion of exogenous variables gives

a spatial-temporal insight of the system, which is helpful to understand the voltage dynamic to be predicted (and controlled).

Since models are obtained in state-space representation, there are good applications that can be implemented to produce a linear control approach. Additionally, these linear state-space representation can be integrated with Kalman filtering approaches that can increase the performance of the prediction by filtering and smoothing the responses. For this, it would be required to explore more carefully the assumption of independent and identically distributed random variables, and normality in the inputs. An alternative could be the integration of EKF and UKF, which consider a non-optimised linear model and non-Gaussian variables.

Future work suggested after this research includes the integration of different exogenous variables that helps on improving the explanation of the obtained linear model (and the assumption in the residuals). Additionally, the proof-of-concept for the implementation of this approach in a control strategy that catch relevant quasi-dynamics in the voltage control problem is suggested to be the next step and validate its effectiveness in a real-time model application.

4.9 Conclusions

In this chapter, the application of the proposed metrics to obtain a data-driven time-series modeling approach is introduced. Section 4.2 presents general considerations of the reference system that will be used to explore the proposed approach. Once the scenario to be explored is detailed, Section 4.3 presents a revision of stationarity conditions in the time-series measured data, which is a condition to apply time-series linear regressions. The proposed methodology is finally introduced in Section 4.4, which becomes one of the main contributions of this work. This is summarized in an algorithm that reviews the current condition of the distribution system using available measurements and builds a linear model. This is applied to an unbalanced distribution system, producing a model that helps predict voltage in each of the phases or available measured nodes of the system. Results over each stage are presented for each measured node/phase.

A first revision of data is introduced in Section 4.4.1, also attempting an initial regression model and a revision of initial assumptions to check if the regression model can explain statistically the voltage quasi-dynamics. Different data-driven modeling approaches were compared for voltage quasi-dynamics in distribution networks. Then, Section 4.4.2 is one of the key steps in the proposed methodology, as it involves a revision of the dataset to improve the results obtained after the first guess. Therefore, it is required to develop a data selection based on crit-

ical scenarios that can characterize the quasi-dynamics to be modeled. Additionally, an analysis of collinearity and distribution shape of regressors is performed, and an analysis of cross-validation and Granger-causality is applied to contrast the response analysis of static and time-variant data, defining relevant lags and reducing the regressors to only selected variables that improve the performance of linear regression. The next step is presented in Section 4.4.3, where the proposed *data-driven* technique can reduce model complexity by system clustering and then being integrated with MISO/MIMO regression methods for the definition of control models. Different model regression techniques were presented and discussed for this specific problem, which showed promising results for the approach based on the Koopman operator (DMDc). The obtained linear *data-driven* representation is helpful to produce state-space linear representation, compatible with model predictive control applications. The initial assumptions were checked again in Section 4.4.4, showing an important improvement compared to the reference initial case. Nevertheless, the initial assumptions were not finally fulfilled in the statistical tests, even if the predicted values were close to the expected values in the training and validation. The previous models were validated using the original data with 1-minute resolution in Section 4.5, and the results showed that the capability of prediction is not highly reduced, considering the difference of granularity in the data measurements. In Section 4.6, it was explored if the obtained models were capable of giving a good voltage prediction beyond one step ahead, even if the previous results showed that it is not possible to conclude anything about their capability of predicting. It is shown that the responses still provide reasonable results after 2-3 steps ahead. The last part of the proposed methodology is explained in Section 4.7, in which an integration of prediction interval based on bootstrapping and cross-validation techniques was explored. This can complement the previous results, and even if the initial assumptions were not fulfilled, it is possible to justify statistically that the results showed validity for the obtained models. The fact that the prediction values remain within the prediction intervals makes the model suitable for short-term prediction.

Chapter 5

Conclusion and Future Work

Data-driven methods aimed at enhancing the integration of renewable energy sources are now recognised as integral components of smart grid applications. These methods offer significant opportunities for electric networks to achieve more efficient and reliable distribution system planning and operation. Voltage control emerges as a crucial function required to upgrade distribution systems into smart grids. The objective is to integrate customers and renewable energy sources by employing a model that aligns with the current situation, without being constrained by generic assumptions of fixed power profiles or network models. In this thesis, novel metrics based on measured data are not only tested but also proposed to characterise the system. These metrics are considered as variables that evolve in space-time, allowing for a comprehensive description of the distribution system without the need for in-depth knowledge of network topology or conventional electrical parameters. Instead, a representation of quasi-dynamics is introduced, integrating the electric system with customer and weather behaviours, while retaining an electric interpretation of the metrics. Moreover, time-series analysis has proven to be an effective tool for modelling (and, it is anticipated after this thesis, for control) based solely on measurable data. This approach aids in describing the current state of the distribution system. Various regression techniques, including autoregressive components and Koopman operators, have been tested to investigate how selected regressors and lags can explain the system under study. The outcome is a reduced-order linear model that captures relevant dynamics, which will prove valuable in the voltage control approach. Statistical analysis has been conducted on the obtained results to comprehend the models and their capabilities, as well as to identify avenues for future research to enhance their performance. This chapter provides a concise overview of the work conducted in this thesis, highlighting the main contributions and discussing potential directions for future

investigations.

5.1 Conclusions and contributions

The main objective of this thesis was to develop an approach for analysing measurable data and generating reduced-order linear models that capture the relevant quasi-dynamics of voltage in distribution systems. This approach takes into account the time-series behaviour of customers and the high penetration of renewable energy sources, with the potential to be integrated with a control approach like MPC. The proposed approach considers operational constraints and uncertainties, such as location, rated power, and variability. The key difference between conventional approaches and the approach adopted in this thesis is the utilisation of measurable data to construct models in a purely data-driven manner. The metrics developed in this thesis are based on the concept of electric distance and perturbation size, providing information on the location and impact within the distribution system. These metrics can be easily integrated into a statistical-based approach, offering a robust representation of the system and establishing a baseline for comparison and improvement through time-series analysis of the obtained regressors (measurable data). Although the results were not fully explained from a statistical perspective (meaning that the initial assumptions for linear regression were not entirely met), the integration of the proposed metrics demonstrated a significant improvement over initial conditions. The predicted values remained within the obtained prediction interval, which is satisfactory for short-term control approaches. The main contributions put forward in this thesis are now listed.

- Chapter 2 primarily focused on conducting a comprehensive review of the state-of-the-art literature to identify the most suitable approach within the research area for developing a model based solely on measurable data. In this chapter, a thorough understanding of the voltage control problem in the context of high-level integration of renewable energy was developed. The fundamentals and traditional assumptions were revisited, leading to the realisation that the problem addressed in this thesis is significantly different from critical points on the conventional power-voltage curve. Instead, the problem can be effectively approximated using a linear representation without compromising precision or understanding of the actual operational conditions. Based on this understanding, the voltage control objective for a potential control approach was introduced, which aims to maximise the integration of renewable energy into the distribution system while adhering

to voltage restrictions. To achieve this objective, it is crucial to have a reliable model that accurately represents the current status of the distribution system. This model can then be integrated into a control approach such as MPC to effectively manage disturbances and exogenous variables that impact voltage operation. The challenges of modelling time-series data and ensuring observability and controllability in distribution systems were discussed. Distribution systems are only partially observable and have limited controllability, presenting a challenge when developing real-time models. Furthermore, these models should account for the quasi-dynamics associated with electric parameters, such as system topology and rated powers for generation and consumption, as well as uncertainties related to exogenous variables, weather conditions, and spatial location of components. To address these challenges, it is necessary to develop new metrics that not only represent the traditional electric parameters but also provide insights into the actual status of the network based solely on measurable data. These metrics should capture the dynamics of the distribution system and enable a comprehensive understanding of its behaviour. Lastly, various data-driven modelling approaches were reviewed, highlighting their applications in voltage control for both conventional and future distribution grids. These approaches offer valuable insights and methodologies for developing effective control strategies based on measurable data, further emphasising the importance of data-driven modelling in addressing the challenges of voltage control in modern distribution systems.

- Chapter 3 focuses on identifying metrics and time-series measurements that aid in describing key features of distribution systems. The chapter begins by investigating the utilisation of measurement data to gain insights into the controllability and observability aspects of voltage in distribution systems. Specifically, it examines information related to nodes that are most affected by power injections or perturbations, as well as an electric distance metric that provides spatial information about voltage variations. A data-driven approach is proposed to enhance the analysis of time-series measured data by characterising relevant inputs and outputs for the modelling process. The chapter explores significant information derived from power fluctuations, voltage covariance, and correlation to improve the understanding of distribution system behaviour. The main contributions of this chapter include the development of two novel metrics, namely M^P and M^Q, which serve as alternatives for identifying and quantifying voltage perturbations based on nodal voltage and injected line power measurements. Additionally, a nor-

malised and averaged covariance matrix of nodal voltages is proposed as a valuable proxy measure for electrical distance, allowing for the classification of node perturbations based on their spatial distance from the source of perturbations. These metrics are evaluated using various types of power/-voltage fluctuations, demonstrating their effectiveness in capturing relevant information from the available data and describing the distribution system. The metrics employed in this study successfully captured the characteristics of the distribution system under various conditions, including the presence of different devices such as capacitor banks and OLTCs. Additionally, the metrics were able to detect whether the system was configured in a radial or meshed topology. This indicates the effectiveness of the metrics in describing and distinguishing different system configurations and the impact of various control elements. Such insights are valuable for understanding the behaviour and performance of the distribution system under different operating conditions. The proposed data-driven approach shows promise in reducing model complexity by providing a criterion for node system clustering, while also capturing the voltage quasi-dynamics. Overall, this chapter provides a reference for characterising key parameters necessary for constructing time-series models without sacrificing the interpretability of the underlying distribution system. Certain portions of this contribution have been published in a conference paper, and a submission to a journal is currently under review (see Appendices A, B, C).

- Chapter 4 of this thesis compares different data-driven modelling approaches to assess their effectiveness in capturing voltage quasi-dynamics in distribution networks. The goal is to generate reduced-order models that are suitable for control applications. The proposed data-driven technique aims to reduce model complexity by establishing a criterion for system clustering and integrating it with MISO/MIMO regression methods. Initially, a review of critical days was conducted to characterise the distribution system under standard conditions. The response analysis of static and time-variant responses was then contrasted for the proposed regressors to determine relevant lags. This involved utilising cross-correlation analysis and Granger-causality analysis. A significant original contribution of this chapter is the introduction of an algorithm that assesses the current condition of the distribution system using available measurements and builds a data-based linear model. One key distinction between this thesis and others in the state-of-the-art is the approach to capturing quasi-dynamics. While traditional system identification techniques consider stochastic inputs as exogenous variables, this thesis treats

them as part of the system with their own quasi-dynamics to be identified. This methodology is compatible with various time-series linear regression approaches, including autoregressive models such as ARX, ARMAX, those based on the principle of Koopman operators or Subspace identification methods. Among these approaches, ARX and DMDc demonstrate the best performances, particularly in producing a state-space linear representation. OKID-ERA was occasionally produce better results, but not consistently and after an exhaustive exploring in the tuning process. Although the results did not fully meet the initial assumptions, the short-term predictions remained within the prediction intervals, which statistically support their acceptance for modelling and control purposes. The corresponding validation of results showed a consistency in the obtained responses, and the models are able to produce three-phases responses for distribution system with unbalance conditions. Certain portions of the contributions presented in this chapter are planned to be submitted for publication in a journal (see Appendix D).

In conclusion, this thesis demonstrates the potential application of time-series data-driven modelling for capturing the quasi-dynamics of a system. The approach integrates the stochastic behaviour of variables used as regressors, making it suitable for plug-and-play real-time applications. The use of measurable data enables insights into the distribution system without relying on prior knowledge of traditional electric system parameters. This representation, based on metrics, allows the model to be adapted over time according to the current system conditions. The obtained results are reinforced by statistical analysis, which helps assess the regressors' capabilities in describing the distribution system.

5.2 Final discussion and future research directions

The results obtained in this thesis demonstrate the effectiveness of the proposed metrics and methodology in describing distribution systems under the influence of different devices, controllers, and topological configurations. However, it is important to note that the impact of measurement noise was not considered in this study, and its inclusion could significantly affect the performance of the results. Furthermore, it should be acknowledged that the availability of comprehensive measurement data, as utilised in this research, may not be readily accessible in practical scenarios. Therefore, it would be necessary to complement this methodology with an analysis to determine the minimum required measurements or key measurement points needed to construct a model that captures the essential quasi-dynamics of the system.

The modelling approach primarily focused on voltage variations and did not encompass other variables such as frequency variations or line loadability. However, it is expected that the approach would not be significantly impacted by the inclusion of these variables. Additionally, constraints such as unbalance limits or component loading were not specifically discussed or evaluated in this study. It would be interesting to investigate the metrics and modelling approach under these constraints to assess their impact and effectiveness in capturing the system behaviour.

In view of the results presented in this thesis, there exists several avenues for possible future work in useful technical directions:

- The results presented in Chapter 3 were obtained using synthetic data generated from simulations, which provided access to nodal and line data for comparisons and understanding the behaviour of the proposed metrics. However, it is worth noting that the proposed metrics do not require full access to nodal voltages for calculation. An important question to be addressed in future research is what information the proposed metrics can provide about the state or behaviour of unobserved parts of the network when measurements are incomplete. This is particularly relevant in the case of a partially observable network with real stochastic behaviour. Future work will focus on identifying the key variables for observation and control, incorporating them into a data-driven model, and utilising them for voltage control. Additionally, exploring different locations for key measurements to ensure that relevant quasi-dynamics are captured in a partially observable system will be investigated. A methodology based on measurable data, such as electric distance or similar parameters, can be developed to determine the optimal position of these measurements. While this thesis has a strong practical focus, there are also works that employ a more system-theoretical background. The research conducted by Professor Van den Hof's group on the identification and identifiability of networked systems, as highlighted in papers such as [298–300], is highly relevant to the problem considered in this thesis. Incorporating their developments can provide a system-theoretic foundation to support the proposed metrics and methodology.

- In Chapter 4, a pure data-driven model was obtained to represent the distribution system without any previous knowledge of electric characteristic. This was done to assess the capabilities of the methodology in extreme scenarios. However, there is also possible to access to other historical/measured data that has been carefully updated for the owner of the model. This meth-

odology could be integrated with previous knowledge of the system to improve the performance of predictions by reducing the amount of information to be explained from regressors. A hybrid data-driven/model approach can be explored and compared with the one obtained in this methodology. Also, the integration of different exogenous variables that helps on improving the explanation of the obtained linear model is a future work (that means, revising other variables that helps on guarantee the initial assumptions in the residuals of the obtained model to be linear). Additionally, the purpose of this thesis was to obtain a linear model that can be used develop (ideally) linear control. That means, it would be ideal to have inputs and residuals that follows the assumptions for linear models. Since models are obtained in state-space representation, they can be integrated with Kalman filtering and smoothing approaches. For this, it would be required to explore the integration of EKF and UKF, which consider a non-optimised linear model and non-Gaussian variables. Finally, it is proposed to develop a proof-of-concept for the implementation of this modelling approach in a control strategy (such as MPC) and validate its effectiveness in a real-time model application.

Acronyms

ACF Autocorrelation Function. xii–xiv, 155, 157–159, 177–181, 213–222

ADF Augmented Dickey-Fuller. 157, 160

AGC Automatic Generation Control. 5

AIC Akaike Information Criterion. 160, 175, 213

ANM Active Network Management. 5, 140

ANN Artificial Neural Network. 41, 137, 214

ARCH Autoregressive Conditional Heteroskedasticity. 187, 188, 743, 757–759

ARIMA AutoRegressive Integrated Moving Average. 151

ARMA Auto-Regressive-Moving-Average Model. 33, 151

ARMAX Auto-Regressive-Moving-Average Models with Exogenous Inputs Model. 42, 44, 49, 51, 137, 142, 161, 168, 169, 174–177, 204, 207–210, 226, 227, 230–240, 246, 254, 693, 723, 746, 747, 755

ARX Auto-Regressive Exogenous. 42, 44, 49, 51, 142, 161, 167–169, 174–177, 204, 207–210, 226, 227, 230–240, 246, 254, 688, 718, 744–746, 754

BCV Blocked Cross-Validation. xiv, 241–245

BIC Bayesian Information Criterion. 160, 175, 213

CDF Cumulative distribution function. 64, 182, 183, 243

CENELEC European Committee for Electrotechnical Standardization. 65

CRA Common Rank Approximation. 48

DAE Differential–Algebraic Equation. 55, 56, 143, 144

DER Distributed Energy Resource. 36, 42, 140

DFIG Doubly-Fed Induction Generator. 42

DG Distributed Generation Unit. 22, 33, 36, 37, 44, 85

DMD Dynamic Mode Decomposition. 43, 49, 51, 142, 162, 163

DMDc Dynamic Mode Decomposition with Control. xiii, xiv, xix, xxii, xxiii, 163, 169, 174–179, 188, 204, 207–210, 213–216, 219, 220, 226, 227, 230–240, 243, 244, 246, 249, 254, 699, 728, 743, 748, 749, 755, 757, 758

ECDF Empirical cumulative distribution function. ix, 65, 66, 69, 72, 243

eDMD Extended DMD. 43

EKF Extended Kalman Filter. 40, 248, 256

EMTP Electromagnetic Transients Program. 38

EPRI Electric Power Research Institute. 58

ERA Eigensystem Realization Algorithm. 42, 50, 51, 165, 167

ESS Energy Storage System. 33

IED Intelligent Electronic Device. 40

IKF Iterated Kalman Filter. 40

IOM Input/Output Model. 40

KS Kolmogorov–Smirnov. 186

LTI Linear time-invariant. 153, 154, 161

MIMO Multiple-Input and Multiple-Output. vii, xiii–xv, xviii, xix, xxiii, 50, 142, 153, 169, 202–204, 209–214, 219–223, 226–229, 231–240, 244, 245, 247, 249, 253, 754, 758, 759

MISO Multiple-Input and Single-Output. vii, xiii–xv, xvii–xix, xxiii, 50, 142, 153, 202–209, 213–218, 223–227, 230–239, 243–245, 247, 249, 253, 744, 757, 758

MPC Model Predictive Control. 32–34, 50, 52, 251, 252, 256

NARMAX Nonlinear Auto-Regressive-Moving-Average Model with Exogenous Inputs Model. 43, 137

NARX Nonlinear Auto-Regressive Model with Exogenous Inputs Model. xiii, xiv, xxiii, 167, 169, 174–177, 180, 181, 188, 204, 207–210, 213, 214, 217, 218, 221, 222, 226, 227, 230–240, 247, 743, 757–759

NN Neural Network. 167

OKID Observed/Kalman Filter Identification. 42, 50, 51, 167

OKID-ERA Observed/Kalman Filter Identification and Eigensystem Realization Algorithm. 164, 169, 174–177, 204, 207–210, 213, 223, 226, 227, 230–240, 246, 254, 700, 729, 749–753, 756

OLTC On-Load Tap Changer. vi, vii, x, 4, 9, 30, 33, 48, 54, 57, 65, 74–77, 86–88, 91, 92, 100, 134, 145, 253, 343, 389, 431, 474, 502, 531, 560, 569, 609, 648

PACF Partial Autocorrelation Function. xii, 156–159

PBM Physics Based Model. 39, 40

PCA Principal Component Analysis. vii, xvi, xxi, xxii, 121, 122, 124–128, 137, 569, 571, 573, 575–577, 579, 581–583, 585, 587, 589, 591, 593, 595–597, 599, 601–603, 605, 607, 609, 611, 613, 615, 617, 619, 621–623, 625, 627, 629, 631, 633, 635, 637, 639, 641–643, 645, 647–649, 651, 653, 655, 657, 659, 661–663, 665, 667–669, 671, 673, 675, 677, 679, 681, 683, 685, 687

PDF Probability density function. 140, 141

PMU Phasor Measurement Unit. 40, 134

PV Photovoltaic. 9, 39, 45, 57, 62, 63, 65, 74, 144, 145, 156

RKHS Reproducing Kernel Hilbert Space. 41

SFA Shifted Frequency Analysis. 38

SINDy Sparse Identification of Nonlinear Dynamics. 43, 167

SISO Single-Input and Single-Output. 50

SVC Static VAR Compensator. 30, 39

SVD Singular Value Decomposition. 163, 164, 166, 169, 174

TSSCV Time-Series Split Cross-Validation. xiv, 241–245

UKF Unscented Kalman Filter. 40, 248, 256

VAR Vector Autoregressive Model. 201

VDE Verband der Elektrotechnik, Elektronik und Informationstechnik. 65

VVO Voltage/Var Optimization. 33

References

[1] G. Simard. IEEE Grid Vision 2050. *IEEE Grid Vision 2050*, pages 1–93, April 2013.

[2] J.M. Gers. *Distribution System Analysis and Automation*. IET Digital Library, UK, 2nd edition edition, 2020.

[3] A. Anwar, N.K. Roy, and H.R. Pota. Voltage stability analysis with optimum size and location based synchronous machine DG. In *AUPEC 2011*, pages 1–5, September 2011.

[4] J.A.D. Massignan, B.R. Pereira, and J.B.A. London. Load Flow Calculation with Voltage Regulators Bidirectional Mode and Distributed Generation. *IEEE Transactions on Power Systems*, 32(2):1576–1577, March 2017.

[5] A.T. Procopiou and L.F. Ochoa. Voltage Control in PV-Rich LV Networks Without Remote Monitoring. *IEEE Transactions on Power Systems*, 32(2):1224–1236, March 2017.

[6] K. Baker, A. Bernstein, C. Zhao, and E. Dall'Anese. Network-cognizant design of decentralized Volt/VAR controllers. In *2017 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, April 2017.

[7] J. Ren, J. Hu, R. Deng, D. Zhang, Y. Zhang, and X.. Shen. Joint Load Scheduling and Voltage Regulation in the Distribution System With Renewable Generators. *IEEE Transactions on Industrial Informatics*, 14(4):1564–1574, April 2018.

[8] A.M. Annaswamy and M. Amin. IEEE Vision for Smart Grid Controls: 2030 and Beyond. *IEEE Vision for Smart Grid Controls: 2030 and Beyond*, pages 1–168, June 2013.

[9] T. Lie and R. Schlueter. Strong local observability and controllability of power systems. In *1991., IEEE International Sympoisum on Circuits and Systems*, pages 970–973 vol.2, June 1991.

[10] L.Y. Wang, F. Lin, and W. Chen. Controllability, Observability, and Integrated State Estimation and Control of Networked Battery Systems. *IEEE Transactions on Control Systems Technology*, 26(5):1699–1710, September 2018.

[11] U. Mhaskar and A. Kulkarni. Power oscillation damping using FACTS devices: modal controllability, observability in local signals, and location of transfer function zeros. *IEEE Transactions on Power Systems*, 21(1):285–294, February 2006.

[12] J. Qi, J. Wang, H. Liu, and A.D. Dimitrovski. Nonlinear Model Reduction in Power Systems by Balancing of Empirical Controllability and Observability Covariances. *IEEE Transactions on Power Systems*, 32(1):114–126, January 2017.

[13] Y. Zhan, X. Xie, and Y. Wang. Impedance Network Model Based Modal Observability and Controllability Analysis for Renewable Integrated Power Systems. *IEEE Transactions on Power Delivery*, 36(4):2025–2034, August 2021.

[14] L.F. Ochoa, C.J. Dent, and G.P. Harrison. Distribution Network Capacity Assessment: Variable DG and Active Networks. *IEEE Transactions on Power Systems*, 25(1):87–95, February 2010.

[15] L.C. Alwan and H.V. Roberts. Time-Series Modeling for Statistical Process Control. *Journal of Business & Economic Statistics*, 6(1):87–95, 1988.

[16] H. Akaike and T. Nakagawa. *Statistical Analysis and Control of Dynamic Systems*. Mathematics and its Applications. Springer Netherlands, 1988.

[17] S. Hagimura, T. Saitoh, and Y. Yagihara. Application of time series analysis and modern control theory to the cement plant. *Annals of the Institute of Statistical Mathematics*, 40(3):419–438, September 1988.

[18] B.F. Crabtree, S.C. Ray, P.M. Schmidt, P.T. O'Connor, and D.D. Schmidt. The individual over time: Time series applications in health care research. *Journal of Clinical Epidemiology*, 43(3):241–260, January 1990.

[19] X.S. Feng. *Dynamic equivalencing of distribution network with embedded generation*. PhD thesis, University of Edimburgh, 2012.

[20] S.M. Zali. *Equivalent dynamic model of distribution network with distributed generation*. PhD thesis, The University of Manchester, 2012.

[21] B. Friedland. *Control System Design: An Introduction to State-Space Methods*. Dover Publications Inc., Mineola, NY, May 2005.

[22] P.W. Sauer, M.A. Pai, and J.H. Chow. *Power System Dynamics and Stability: With Synchrophasor Measurement and Power System Toolbox*. Wiley-IEEE Press, Hoboken, NJ, USA, 2nd edition, September 2017.

[23] X. Qin, X. Shen, H. Sun, and Q. Guo. A Quasi-Dynamic Model and Corresponding Calculation Method for Integrated Energy System with Electricity and Heat. *Energy Procedia*, 158:6413–6418, February 2019.

[24] D. Raoofsheibani, P. Hinkel, and W.H. Wellssow. A Quasi-Dynamic Tool for Validation of Power System Restoration Strategies at Distribution Level. In *2019 IEEE Milan PowerTech*, pages 1–6, June 2019.

[25] I. Dobson and H.D. Chiang. Towards a theory of voltage collapse in electric power systems. *Systems & Control Letters*, 13(3):253–262, September 1989.

[26] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T.V. Cutsem, and V. Vittal. Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. *IEEE Transactions on Power Systems*, 19(3):1387–1401, August 2004.

[27] A. Chakrabarti, D.P. Kothari, m.A. K, and A. De. *An Introduction to Reactive Power Control and Voltage Stability in Power Transmission Systems*. PHI Learning Pvt. Ltd., January 2010.

[28] J. Machowski, J.W. Bialek, and J.R. Bumby. *Power System Dynamics: Stability and Control, 2nd Edition*. Wiley, 2nd edition, 2008.

[29] J. Hossain and H.R. Pota. *Robust Control for Grid Voltage Stability: High Penetration of Renewable Energy: Interfacing Conventional and Renewable Power Generation Resources*. Power Systems. Springer Singapore, 2014.

[30] R. Arghandeh and Y. Zhou. *Big data application in power systems*. Elsevier, Amsterdam, Netherlands; Kidlington, Oxford, 2018. OCLC: 1013594912.

[31] S.L. Brunton and J.N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, February 2019.

[32] J. Dong, Y. Xue, M. Olama, T. Kuruganti, J. Nutaro, and C. Winstead. Distribution Voltage Control: Current Status and Future Trends. In *2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, pages 1–7, June 2018. ISSN: 2329-5767.

[33] X. Xu, K. Li, F. Qi, H. Jia, and J. Deng. Identification of microturbine model for long-term dynamic analysis of distribution networks. *Applied Energy*, 192:305–314, April 2017.

[34] X. Feng, Z. Lubosny, and J.W. Bialek. Identification based Dynamic Equivalencing. In *2007 IEEE Lausanne Power Tech*, pages 267–272, July 2007.

[35] S. Chanda and B. Das. Identification of weak buses in a power network using novel voltage stability indicator in radial distribution system. In *India International Conference on Power Electronics 2010 (IICPE2010)*, pages 1–4, January 2011.

[36] C. Viggiano, P. Trodden, E. Caicedo, and W. Alfonso. Data-Driven Characterisation of Distribution Systems for Modelling and Control Applications. In *2022 International Conference on Smart Energy Systems and Technologies (SEST)*, pages 1–6, September 2022.

[37] C. Viggiano, P. Trodden, E. Caicedo, and W. Alfonso. Distribution Systems Modelling by Data-Driven Voltage Characterisation for Control Applications-Part I: Input Analysis. *Submitted to IEEE Transactions on Power Systems*, pages 1–8, 2022.

[38] C. Viggiano, P. Trodden, E. Caicedo, and W. Alfonso. Distribution Systems Modelling by Data-Driven Voltage Characterisation for Control Applications-Part II: Case Studies. *Submitted to IEEE Transactions on Power Systems*, pages 1–10, 2022.

[39] C. Viggiano, P. Trodden, E. Caicedo, and W. Alfonso. Data-Driven Time-Series-based approach for modelling of distribution system with high penetration of renewable energy sources. *To be submitted*, 2023.

[40] J.W. Simpson-Porco, F. Dörfler, and F. Bullo. Voltage collapse in complex power grids. *Nature Communications*, 7:10790, February 2016.

[41] CIGRÉ WG 32.03. Tentative classification and terminologies relating to stability problems of power systems. Technical report, Conseil International des Grands Réseaux Électriques, CIGRÉ, 1978.

[42] IEEE, Task Force on Terms & Definitions, System Dynamic Performance Subcommittee, Power System Engineering Committee. Proposed Terms amp; Definitions for Power System Stability. *IEEE Transactions on Power Apparatus and Systems*, PAS-101(7):1894–1898, July 1982.

[43] I. Dobson, T. Van Cutsem, C. Vournas, C. Demarco, M. Venkatasubramanian, T. Overbye, and C. Canizares. Voltage Stability Assessment: Concepts, Practices and Tools. *IEEE Power Engineering Society, Power System Stability Subcommittee Special Publication*, 11:21–22, January 2002.

[44] P. Kundur. *Power System Stability and Control*. McGraw-Hill Education, January 1994.

[45] M. Ghaffarianfar and A. Hajizadeh. Voltage Stability of Low-Voltage Distribution Grid with High Penetration of Photovoltaic Power Units. *Energies*, 11 (8):1960, August 2018.

[46] A. Wiszniewski. New Criteria of Voltage Stability Margin for the Purpose of Load Shedding. *IEEE Transactions on Power Delivery*, 22(3):1367–1371, July 2007.

[47] L.F. Ochoa, A. Padilha-Feltrin, and G.P. Harrison. Evaluating distributed generation impacts with a multiobjective index. *IEEE Transactions on Power Delivery*, 21(3):1452–1458, July 2006.

[48] N.K. Roy. *Voltage Stability Enhancement of Distribution Systems with Renewable Energy*. Doctoral Thesis, The University of New South Wales, Australia, 2013.

[49] M. Alonso and H. Amaris. Voltage stability in distribution networks with DG. In *2009 IEEE Bucharest PowerTech*, pages 1–6, June 2009.

[50] C.W. Taylor. *Power system voltage stability*. McGraw-Hill Ryerson, Limited, 1994.

[51] S.A.A. Kazmi, M.K. Shahzaad, and D.R. Shin. Voltage Stability Index for Distribution Network connected in Loop Configuration. *IETE Journal of Research*, 63(2):281–293, March 2017.

[52] R. Al-Abri. *Voltage Stability Analysis with High Distributed Generation (DG) Penetration*. PhD thesis, University of Waterloo, Canada, August 2012.

[53] N. Nikpour. *Dynamic and static voltage stability analysis of distribution systems in the presence of distributed generation*. PhD thesis, University of British Columbia, 2016.

[54] L. Aolaritei, S. Bolognani, and F. Dörfler. A distributed voltage stability margin for power distribution networks. *IFAC-PapersOnLine*, 50(1):13240–13245, July 2017.

[55] Y. Li, X. Tian, C. Liu, Y. Su, L. Li, L. Zhang, Y. Sun, and J. Li. Study on voltage control in distribution network with renewable energy integration. In *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pages 1–5, November 2017.

[56] L. Aolaritei, S. Bolognani, and F. Dörfler. Hierarchical and Distributed Monitoring of Voltage Stability in Distribution Networks. *IEEE Transactions on Power Systems*, 33(6):6705–6714, November 2018.

[57] M. Todescato, J.W. Simpson-Porco, F. Dörfler, R. Carli, and F. Bullo. Voltage stress minimization by optimal reactive power control. *arXiv:1602.01969 [math.OC]*, February 2016.

[58] D.P. Nedic, I. Dobson, D.S. Kirschen, B.A. Carreras, and V.E. Lynch. Criticality in a cascading failure blackout model. *International Journal of Electrical Power & Energy Systems*, 28(9):627–633, November 2006.

[59] I. Dobson, B.A. Carreras, V.E. Lynch, and D.E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2): 026103, June 2007.

[60] R. Pfitzner, K. Turitsyn, and M. Chertkov. Statistical classification of cascading failures in power grids. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–8, July 2011.

[61] E. Cotilla-Sanchez, P.D.H. Hines, and C.M. Danforth. Predicting Critical Transitions From Time Series Synchrophasor Data. *IEEE Transactions on Smart Grid*, 3(4):1832–1840, December 2012.

[62] D. Podolsky and K. Turitsyn. Critical slowing-down as indicator of approach to the loss of stability. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 19–24, November 2014.

[63] G. Ghanavati, P.D.H. Hines, T.I. Lakoba, and E. Cotilla-Sanchez. Understanding Early Indicators of Critical Transitions in Power Systems From Autocorrelation Functions. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 61(9):2747–2760, September 2014.

[64] G. Ghanavati, P.D.H. Hines, and T.I. Lakoba. Identifying Useful Statistical Indicators of Proximity to Instability in Stochastic Power Systems. *IEEE Transactions on Power Systems*, 31(2):1360–1368, March 2016.

[65] S. Kéfi, V. Dakos, M. Scheffer, E.H.V. Nes, and M. Rietkerk. Early warning signals also precede non-catastrophic transitions. *Oikos*, 122(5):641–648, 2013.

[66] P. Jorgensen, J.S. Christensen, and J.O. Tande. Probabilistic load flow calculation using Monte Carlo techniques for distribution network with wind turbines. In *8th International Conference on Harmonics and Quality of Power. Proceedings (Cat. No.98EX227)*, volume 2, pages 1146–1151 vol.2, October 1998.

[67] T.R. Ricciardi, K. Petrou, J.F. Franco, and L.F. Ochoa. Defining Customer Export Limits in PV-Rich Low Voltage Networks. *IEEE Transactions on Power Systems*, 34(1):87–97, January 2019.

[68] Andreas Procopiou. *Active Management of PV-Rich Low Voltage Networks*. PhD thesis, University of Manchester, 2017.

[69] A.T. Procopiou, K. Petrou, L.F. Ochoa, T. Langstaff, and J. Theunissen. Adaptive Decentralized Control of Residential Storage in PV-Rich MV–LV Networks. *IEEE Transactions on Power Systems*, 34(3):2378–2389, May 2019.

[70] A.T. Procopiou, C. Long, and L.F. Ochoa. On the effects of monitoring and control settings on voltage control in PV-rich LV networks. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015.

[71] C. Long, A.T. Procopiou, L.F. Ochoa, G. Bryson, and D. Randles. Performance of OLTC-based control strategies for LV networks with photovoltaics. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015.

[72] K. Petrou, L.F. Ochoa, A.T. Procopiou, J. Theunissen, J. Bridge, T. Langstaff, and K. Lintern. Limitations of Residential Storage in PV-Rich Distribution Networks: An Australian Case Study. In *2018 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2018.

[73] J.F. Franco, A.T. Procopiou, J. Quirós-Tortós, and L.F. Ochoa. Advanced control of OLTC-enabled LV networks with PV systems and EVs. *IET Generation, Transmission & Distribution*, 13(14):2967–2975, 2019.

[74] P.N. Vovos, A.E. Kiprakis, A.R. Wallace, and G.P. Harrison. Centralized and Distributed Voltage Control: Impact on Distributed Generation Penetration. *IEEE Transactions on Power Systems*, 22(1):476–483, February 2007.

[75] J.W. Simpson-Porco, F. Dörfler, and F. Bullo. Voltage Stabilization in Microgrids via Quadratic Droop Control. *IEEE Transactions on Automatic Control*, 62(3):1239–1253, March 2017.

[76] I. Hiskens and B. Gong. Voltage Stability Enhancement Via Model Predictive Control of Load. *Intelligent Automation & Soft Computing*, 12, January 2006.

[77] A. Kechroud, J.F.C. Flores, and W.L. Kling. Voltage control in distribution networks using fast model predictive control. In *IEEE PES General Meeting*, pages 1–5, July 2010.

[78] A. Kechroud, J.F.C. Flores, and W.L. Kling. Multiple Models adaptive voltage control in distribution networks. In *45th International Universities Power Engineering Conference UPEC2010*, pages 1–5, August 2010.

[79] Z. Wang, J. Wang, B. Chen, M.M. Begovic, and Y. He. MPC-Based Voltage/Var Optimization for Distribution Circuits With Distributed Generators and Exponential Load Models. *IEEE Transactions on Smart Grid*, 5(5): 2412–2420, September 2014.

[80] M. Armendariz, D. Babazadeh, D. Brodén, and L. Nordström. Strategies to improve the voltage quality in active low-voltage distribution networks using DSO's assets. *Transmission Distribution IET Generation*, 11(1):73–81, 2017.

[81] D. Zarrilli, A. Giannitrapani, S. Paoletti, and A. Vicino. Energy Storage Operation for Voltage Control in Distribution Networks: A Receding Horizon Approach. *IEEE Transactions on Control Systems Technology*, 26(2):599–609, March 2018.

[82] Y. Guo, Q. Wu, H. Gao, X. Chen, J. Ostergaard, and H. Xin. MPC-Based Coordinated Voltage Regulation for Distribution Networks With Distributed Generation and Energy Storage System. *IEEE Transactions on Sustainable Energy*, pages 1–1, 2018.

[83] Y. Guo, Q. Wu, H. Gao, S. Huang, B. Zhou, and C. Li. Double-Time-Scale Coordinated Voltage Control in Active Distribution Networks Based on MPC. *IEEE Transactions on Sustainable Energy*, pages 1–1, 2019.

[84] S. Riverso and G. Ferrari-Trecate. Hycon2 Benchmark: Power Network System. *arXiv:1207.2000 [cs.SY]*, July 2012.

[85] S. Riverso, M. Farina, and G. Ferrari-Trecate. Plug-and-Play decentralized Model Predictive Control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 4193–4198, December 2012.

[86] S. Riverso, M. Farina, and G. Ferrari-Trecate. Plug-and-Play Decentralized Model Predictive Control for Linear Systems. *IEEE Transactions on Automatic Control*, 58(10):2608–2614, October 2013.

[87] L. Huang, J. Coulson, J. Lygeros, and F. Dorfler. Data-Enabled Predictive Control for Grid-Connected Power Converters. *arXiv:1903.07339 [cs.SY]*, March 2019.

[88] Z.A. Khan and D. Jayaweera. Smart Meter Data Based Load Forecasting and Demand Side Management in Distribution Networks With Embedded PV Systems. *IEEE Access*, 8:2631–2644, 2020.

[89] F. Bai, Y. Liu, K. Sun, N. Bhatt, A.D. Rosso, E. Farantatos, and X. Wang. Input signals selection for measurement-based power system ARX dynamic model response estimation. In *2014 IEEE PES T&D Conference and Exposition*, pages 1–7, April 2014.

[90] C. Li, Y. Yu, J. Yan, and Y. Liu. Power system dynamics equivalent model based on AutoRegressive model with eXogenous inputs. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 1947–1952, November 2017.

[91] Y. Yuan, S.H. Low, O. Ardakanian, and C.J. Tomlin. Inverse Power Flow Problem. *IEEE Transactions on Control of Network Systems (Early Access)*, pages 1–12, 2022.

[92] O. Ardakanian, Y. Yuan, R. Dobbe, A. von Meier, S. Low, and C. Tomlin. Event Detection and Localization in Distribution Grids with Phasor Measurement Units. *arXiv: 1611.04653 [cs.SY]*, November 2016.

[93] O. Ardakanian, V.W.S. Wong, R. Dobbe, S.H. Low, A.v. Meier, C.J. Tomlin, and Y. Yuan. On Identification of Distribution Grids. *IEEE Transactions on Control of Network Systems*, 6(3):950–960, September 2019.

[94] A. Primadianto and C. Lu. A Review on Distribution System State Estimation. *IEEE Transactions on Power Systems*, 32(5):3875–3883, September 2017.

[95] C. Carquex, C. Rosenberg, and K. Bhattacharya. State Estimation in Power Distribution Systems Based on Ensemble Kalman Filtering. *IEEE Transactions on Power Systems*, 33(6):6600–6610, November 2018.

[96] T. Sadamoto, A. Chakrabortty, T. Ishizaki, and J. Imura. Dynamic Modeling, Stability, and Control of Power Systems With Distributed Energy Resources: Handling Faults Using Two Control Methods in Tandem. *IEEE Control Systems Magazine*, 39(2):34–65, April 2019.

[97] M. Netto and L. Mili. A Robust Data-Driven Koopman Kalman Filter for Power Systems Dynamic State Estimation. *IEEE Transactions on Power Systems*, 33(6):7228–7237, November 2018.

[98] G. Cavraro and V. Kekatos. Inverter Probing for Power Distribution Network Topology Processing. *IEEE Transactions on Control of Network Systems*, 6(3): 980–992, September 2019.

[99] J.W. Pierre, D. Trudnowski, M. Donnelly, N. Zhou, F.K. Tuffner, and L. Dosiek. Overview of System Identification for Power Systems from Measured Responses. *IFAC Proceedings Volumes*, 45(16):989–1000, July 2012.

[100] A.S. Bretas, A. Rossoni, R.D. Trevizan, and N.G. Bretas. Distribution networks nontechnical power loss estimation: A hybrid data-driven physics model-based framework. *Electric Power Systems Research*, 186:106397, September 2020.

[101] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang. Outage Detection in Partially Observable Distribution Systems Using Smart Meters and Generative Adversarial Networks. *IEEE Transactions on Smart Grid*, 11(6):5418–5430, November 2020.

[102] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, 1999.

[103] M. Gevers. Identification for Control: From the Early Achievements to the Revival of Experiment Design. *European Journal of Control*, 11(4):335–352, January 2005.

[104] L. Ljung. *Perspectives on System Identification*. Linköping University Electronic Press, 2010.

[105] L. Fan, Z. Miao, S. Shah, P. Koralewicz, V. Gevorgian, and J. Fu. Data-Driven Dynamic Modeling in Power Systems: A Fresh Look on Inverter-Based Resource Modeling. *IEEE Power and Energy Magazine*, 20(3):64–76, May 2022.

[106] F. Znidi, H. Davarikia, K. Iqbal, and M. Barati. Multi-layer spectral clustering approach to intentional islanding in bulk power systems. *Journal of Modern Power Systems and Clean Energy*, 7(5):1044–1055, September 2019.

[107] P. Lagonotte. The different electric distances. In *Proceedings of the Tenth Power Systems Computation Conference*. August 1990.

[108] S. Bolognani. Grid Topology Identification via Distributed Statistical Hypothesis Testing. In R. Arghandeh and Y. Zhou, editors, *Big Data Application in Power Systems*, pages 281–301. Elsevier, January 2018.

[109] Y. Liao, Y. Weng, G. Liu, Z. Zhao, C.W. Tan, and R. Rajagopal. Unbalanced multi-phase distribution grid topology estimation and bus phase identification. *IET Smart Grid*, 2(4):557–570, 2019.

[110] Y. Song, D.J. Hill, and T. Liu. Static Voltage Stability Analysis of Distribution Systems Based on Network-Load Admittance Ratio. *IEEE Transactions on Power Systems*, 34(3):2270–2280, May 2019.

[111] M.J. Hossain, H.R. Pota, and V. Ugrinovskii. Short and Long-Term Dynamic Voltage Instability. *IFAC Proceedings Volumes*, 41(2):9392–9397, January 2008.

[112] M.J. Hossain, H.R. Pota, and R.A. Ramos. Improved low-voltage-ride-through capability of fixed-speed wind turbines using decentralised control of STATCOM with energy storage system. *IET Generation, Transmission & Distribution*, 6(8):719–730, August 2012.

[113] M.J. Hossain, H.R. Pota, M.A. Mahmud, and R.A. Ramos. Investigation of the Impacts of Large-Scale Wind Power Penetration on the Angle and Voltage Stability of Power Systems. *IEEE Systems Journal*, 6(1):76–84, March 2012.

[114] S. Henschel. *Analysis of electromagnetic and electromechanical power system transients with dynamic phasors*. PhD thesis, University of British Columbia, 1999.

[115] A.M. Stankovic, S.R. Sanders, and T. Aydin. Dynamic phasors in modeling and analysis of unbalanced polyphase AC machines. *IEEE Transactions on Energy Conversion*, 17(1):107–113, March 2002.

[116] P.C. Stefanov and A.M. Stankovic. Modeling of UPFC operation under unbalanced conditions with dynamic phasors. *IEEE Transactions on Power Systems*, 17(2):395–403, May 2002.

[117] T. Demiray. *Simulation of Power System Dynamics using Dynamic Phasor Models*. Doctoral Thesis, ETH Zurich, 2008.

[118] T. Yang, S. Bozhko, and G. Asher. Multi-generator system modelling based on dynamic phasor concept. In *2013 15th European Conference on Power Electronics and Applications (EPE)*, pages 1–10, September 2013.

[119] J.R. Martí, H.W. Dommel, B.D. Bonatto, and A.F.R. Barrete. Shifted Frequency Analysis (SFA) concepts for EMTP modelling and simulation of Power System Dynamics. In *2014 Power Systems Computation Conference*, pages 1–8, August 2014.

[120] M.J. Gorman and S. Civanlar. Load/voltage modeling of distribution systems; a system identification approach. In *Proceedings. IEEE Energy and Information Technologies in the Southeast'*, pages 386–389 vol.1, April 1989.

[121] Z. Staroszczyk. Problems in real-time wide band identification of power systems. In *IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference*, volume 2, pages 779–784 vol.2, May 1998.

[122] G. Valverde and V. Terzija. Unscented Kalman filter for power system dynamic state estimation. *IET Generation Transmission & Distribution*, 5(1):29–37, January 2011.

[123] C. Hernández and P. Maya-Ortiz. Comparison between WLS and Kalman Filter method for power system static state estimation. In *2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST)*, pages 47–52, September 2015.

[124] T. Sarkar, A. Rakhlin, and M.A. Dahleh. Nonparametric System identification of Stochastic Switched Linear Systems. *arXiv*, September 2019.

[125] L. Cupelli, A. Esteban, F. Ponci, and A. Monti. Kernel-based online learning for real-time voltage control in distribution networks. *IET Smart Grid*, April 2020.

[126] L. Cupelli, M. Cupelli, F. Ponci, and A. Monti. Data-Driven Adaptive Control for Distributed Energy Resources. *IEEE Transactions on Sustainable Energy*, 10 (3):1575–1584, July 2019.

[127] X. Feng, Z. Lubosny, and J. Bialek. Dynamic Equivalencing of Distribution Network with High Penetration of Distributed Generation. In *Proceedings of the 41st International Universities Power Engineering Conference*, volume 2, pages 467–471, September 2006.

[128] D. Jia, W. Sheng, X. Song, and X. Meng. A system identification method for smart distribution grid. In *2014 International Conference on Power System Technology*, pages 14–19, October 2014.

[129] F.O. Resende, J. Matevosyan, and J.V. Milanovic. Application of dynamic equivalence techniques to derive aggregated models of active distribution network cells and microgrids. In *2013 IEEE Grenoble Conference*, pages 1–6, June 2013.

[130] M. Norgaard, O. Ravn, N.K. Poulsen, and L.K. Hansen. *Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag, London, 2000.

[131] B. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, February 1981.

[132] J.N. Juang and R.S. Pappa. An eigensystem realization algorithm for modal parameter identification and model reduction. *Journal of Guidance, Control, and Dynamics*, 8(5):620–627, September 1985.

[133] Z. Ma, S. Ahuja, and C.W. Rowley. Reduced-order models for control of fluids using the eigensystem realization algorithm. *Theoretical and Computational Fluid Dynamics*, 25(1):233–247, June 2011.

[134] H. Liu, L. Zhu, Z. Pan, J. Guo, J. Chai, Wenpeng Yu, and Y. Liu. Comparison of MIMO system identification methods for electromechanical oscillation damping estimation. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5, July 2016.

[135] M.Q. Phan and R.W. Longman. Relationship Between State-space And Input-output Models Via Observer Markov Parameters. *WIT Transactions on The Built Environment*, 22:16, 1996.

[136] B.O. Koopman. Hamiltonian Systems and Transformation in Hilbert Space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, May 1931.

[137] I. Mezić and A. Banaszuk. Comparison of systems with complex behavior. *Physica D: Nonlinear Phenomena*, 197(1):101–133, October 2004.

[138] J.N. Kutz, S.L. Brunton, B.W. Brunton, and J.L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Society for Industrial and Applied Mathematics, Philadelphia, November 2016.

[139] P.J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, August 2010.

[140] J.H. Tu, C.W. Rowley, D.M. Luchtenburg, S.L. Brunton, and J.N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391, 2014.

[141] M.O. Williams, I.G. Kevrekidis, and C.W. Rowley. A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, December 2015.

[142] M. Netto, Y. Susuki, and L. Mili. Data-Driven Participation Factors for Nonlinear Systems Based on Koopman Mode Decomposition. *IEEE Control Systems Letters*, 3(1):198–203, January 2019.

[143] S.L. Brunton, J.L. Proctor, and J.N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016.

[144] S.A. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. John Wiley & Sons, Sheffield, UK, 1 edition, September 2013.

[145] Y. Susuki and I. Mezić. Nonlinear Koopman Modes and Power System Stability Assessment Without Models. *IEEE Transactions on Power Systems*, 29 (2):899–907, March 2014.

[146] Y. Susuki, I. Mezic, F. Raak, and T. Hikihara. Applied Koopman Operator Theory for Power Systems Technology. *Nonlinear Theory and Its Applications, IEICE*, 7(4):430–459, 2016. arXiv: 1706.00159.

[147] Y. Susuki and K. Sako. Data-Based Voltage Analysis of Power Systems via Delay Embedding and Extended Dynamic Mode Decomposition. *IFAC-PapersOnLine*, 51(28):221–226, January 2018.

[148] M. Netto, V. Krishnan, L. Mili, Y. Susuki, and Y. Zhang. A Hybrid Framework Combining Model-Based and Data-Driven Methods for Hierarchical

Decentralized Robust Dynamic State Estimation. In *2019 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2019. ISSN: 1944-9933.

[149] M. Zhou and J. Buongiorno. Space-Time Modeling of Timber Prices. *Journal of Agricultural and Resource Economics*, 31(1):40–56, 2006. Publisher: Western Agricultural Economics Association.

[150] R.L. Martin and J.E. Oeppen. The Identification of Regional Forecasting Models Using Space: Time Correlation Functions. *Transactions of the Institute of British Geographers*, 66:95–118, 1975. Publisher: [Royal Geographical Society (with the Institute of British Geographers), Wiley].

[151] C. Mocenni, A. Facchini, and A. Vicino. Identifying the dynamics of complex spatio-temporal systems by spatial recurrence properties. *Proceedings of the National Academy of Sciences*, 107(18):8097–8102, May 2010. Publisher: Proceedings of the National Academy of Sciences.

[152] P. Aram, V. Kadirkamanathan, and S.R. Anderson. Spatiotemporal System Identification With Continuous Spatial Maps and Sparse Estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 26(11):2978–2983, November 2015. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[153] C. Fonseca, A. Martins, L. Pereira, and H. Ferreira. Measuring dependence of a space-time ARMAX storage model. In *METMA V - International Workshop on Spatio-Temporal Modelling*, Santiago de Compostela, España, June 2010.

[154] N. Voulis, M. Warnier, and F.M.T. Brazier. Understanding spatio-temporal electricity demand at different urban scales: A data-driven approach. *Applied Energy*, 230:1157–1171, November 2018.

[155] L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, G.C. Yalcin, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. Open database analysis of scaling and spatio-temporal properties of power grid frequencies. *Nature Communications*, 11(1):6362, December 2020. Number: 1 Publisher: Nature Publishing Group.

[156] H. Ma, X. Lei, Z. Li, S. Yu, B. Liu, and X. Dong. Deep-learning based Power System Events Detection Technology Using Spatio-temporal and Frequency Information. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, pages 1–1, 2023. Conference Name: IEEE Journal on Emerging and Selected Topics in Circuits and Systems.

[157] D. Horak, A. Hainoun, G. Neugebauer, and G. Stoeglehner. A review of spatio-temporal urban energy system modeling for urban decarbonization strategy formulation. *Renewable and Sustainable Energy Reviews*, 162:112426, July 2022.

[158] X. Bai, D. Liu, J. Tan, H. Yang, and H. Zheng. Dynamic Identification of Critical Nodes and Regions in Power Grid Based on Spatio-Temporal Attribute Fusion of Voltage Trajectory. *Energies*, 12:780, February 2019.

[159] J. Gulliver and D.J. Briggs. Time–space modeling of journey-time exposure to traffic-related air pollution using GIS. *Environmental Research*, 97(1):10–25, January 2005.

[160] G.P. Harrison and A.R. Wallace. Optimal power flow evaluation of distribution network capacity for the connection of distributed generation. *IEE Proceedings - Generation, Transmission and Distribution*, 152(1):115–122, January 2005.

[161] T. Boehme, A.R. Wallace, and G.P. Harrison. Applying Time Series to Power Flow Analysis in Networks With High Wind Penetration. *IEEE Transactions on Power Systems*, 22(3):951–957, August 2007.

[162] G. Gross and F. Galiana. Short-term load forecasting. *Proceedings of the IEEE*, 75(12):1558–1573, December 1987.

[163] B. Wang, N. Duan, and K. Sun. A Time–Power Series-Based Semi-Analytical Approach for Power System Simulation. *IEEE Transactions on Power Systems*, 34(2):841–851, March 2019.

[164] X. Jiang, Y.C. Chen, and A.D. Domínguez-García. A set-theoretic framework to assess the impact of variable generation on the power flow. *IEEE Transactions on Power Systems*, 28(2):855–867, May 2013.

[165] J. Nazarko and Z. Styczynski. Application of statistical and neural approaches to the daily load profiles modelling in power distribution systems. In *1999 IEEE Transmission and Distribution Conference*, volume 1, pages 320–325 vol.1, April 1999.

[166] D. Hill, D. McMillan, K. Bell, D. Infield, and G. Ault. Application of statistical wind models for system impacts. In *2009 44th International Universities Power Engineering Conference (UPEC)*, pages 1–5, September 2009.

[167] G. Liu and J. Fan. Framework for statistical analysis of homogeneous multicore power grid networks. In *2009 IEEE 8th International Conference on ASIC*, pages 423–426, October 2009.

[168] F. Rassaei, W.S. Soh, and K.C. Chua. A Statistical modelling and analysis of residential electric vehicles' charging demand in smart grids. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, February 2015.

[169] M. Sun, I. Konstantelos, S. Tindemans, and G. Strbac. Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems. In *2016 Power Systems Computation Conference (PSCC)*, pages 1–8, June 2016.

[170] A. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

[171] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[172] G. Cavraro, R. Arghandeh, K. Poolla, and A. von Meier. Data-driven approach for distribution network topology detection. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015.

[173] D. Deka, S. Backhaus, and M. Chertkov. Structure Learning in Power Distribution Networks. *IEEE Transactions on Control of Network Systems*, 5(3): 1061–1074, September 2018.

[174] A. Selim, M. Abdel-Akher, M.M. Aly, S. Kamel, and T. Senjyu. Fast quasi-static time-series analysis and reactive power control of unbalanced distribution systems. *International Transactions on Electrical Energy Systems*, 29(1): 1–14, 2019.

[175] R. Yao, S. Huang, K. Sun, F. Liu, X. Zhang, and S. Mei. A Multi-Timescale Quasi-Dynamic Model for Simulation of Cascading Outages. *IEEE Transactions on Power Systems*, 31(4):3189–3201, July 2016. Conference Name: IEEE Transactions on Power Systems.

[176] M.J. Reno and R.J. Broderick. Predetermined time-step solver for rapid quasi-static time series (QSTS) of distribution systems. In *2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, April 2017. ISSN: 2472-8152.

[177] G. Judah. Power Systems Modelling of a Community Energy Project. In *CIRED Workshop*, Ljubljana, Slovenia, June 2018. Accepted: 2019-12-19T18:20:02Z ISSN: 2032-9628.

[178] W. Niederhuemer and R. Schwalbe. Increasing PV hosting capacity in LV grids with a probabilistic planning approach. In *2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST)*, pages 537–540, September 2015.

[179] M. Milligan, P. Donohoo, and M. O 'malley. Stochastic Methods for Planning and Operating Power System with Large Amounts of Wind and Solar Power. In *11th Annual International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plants Conference*, Lisbon, Portugal, November 2012.

[180] Z. Wang and F.L. Alvarado. Interval arithmetic in power flow analysis. *IEEE Transactions on Power Systems*, 7(3):1341–1349, August 1992.

[181] A.T. Saric and A.M. Stankovic. Model uncertainty in security assessment of power systems. *IEEE Transactions on Power Systems*, 20(3):1398–1407, August 2005.

[182] B. Borkowska. Probabilistic Load Flow. *IEEE Transactions on Power Apparatus and Systems*, PAS-93(3):752–759, May 1974.

[183] R.N. Allan, A.m.L.D. Silva, and R.C. Burchett. Evaluation Methods and Accuracy in Probabilistic Load Flow Solutions. *IEEE Transactions on Power Apparatus and Systems*, PAS-100(5):2539–2546, May 1981.

[184] A.M.L.d. Silva and V.L. Arienti. Probabilistic load flow by a multilinear simulation algorithm. *Transmission and Distribution IEE Proceedings C - Generation*, 137(4):276–282, July 1990.

[185] R. Allan and R. Billinton. Probabilistic assessment of power systems. *Proceedings of the IEEE*, 88(2):140–162, February 2000.

[186] Pei Zhang and S.T. Lee. Probabilistic load flow computation using the method of combined cumulants and Gram-Charlier expansion. *IEEE Transactions on Power Systems*, 19(1):676–682, February 2004.

[187] H. Yu, C.Y. Chung, K.P. Wong, H.W. Lee, and J.H. Zhang. Probabilistic Load Flow Evaluation With Hybrid Latin Hypercube Sampling and Cholesky Decomposition. *IEEE Transactions on Power Systems*, 24(2):661–667, May 2009.

[188] J. Usaola. Probabilistic load flow in systems with wind generation. *Transmission Distribution IET Generation*, 3(12):1031–1041, December 2009.

[189] J. Schwippe, O. Krause, and C. Rehtanz. Probabilistic Load Flow Calculation based on an enhanced convolution technique. In *2009 IEEE PowerTech*, pages 1–6, Bucharest, Romania, June 2009.

[190] J. Schwippe, O. Krause, and C. Rehtanz. Extension of a probabilistic load flow calculation based on an enhanced convolution technique. In *2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE)*, pages 1–6, September 2009.

[191] O. Krause, J. Schwippe, and M. Eghbal. Probabilistic calculus in power system analysis and design. In *AUPEC 2011*, pages 1–6, September 2011.

[192] O. Krause, J. Schwippe, S. Lehnhoff, and C. Rehtanz. Analytic solution of the classic probabilistic load flow problem on a full AC model. In *2011 IEEE PES Innovative Smart Grid Technologies*, pages 1–8, November 2011.

[193] Y. Yuan, J. Zhou, P. Ju, and J. Feuchtwang. Probabilistic load flow computation of a power system containing wind farms using the method of combined cumulants and Gram-Charlier expansion. *IET Renewable Power Generation*, 5(6):448–454, November 2011.

[194] M. Fan, V. Vittal, G.T. Heydt, and R. Ayyanar. Probabilistic Power Flow Studies for Transmission Systems With Photovoltaic Generation Using Cumulants. *IEEE Transactions on Power Systems*, 27(4):2251–2261, November 2012.

[195] E. Janecek and D. Georgiev. Probabilistic Extension of the Backward/Forward Load Flow Analysis Method. *IEEE Transactions on Power Systems*, 27 (2):695–704, May 2012.

[196] Y. Chen, J. Wen, and S. Cheng. Probabilistic Load Flow Method Based on Nataf Transformation and Latin Hypercube Sampling. *IEEE Transactions on Sustainable Energy*, 4(2):294–301, April 2013.

[197] M. Hajian, W.D. Rosehart, and H. Zareipour. Probabilistic Power Flow by Monte Carlo Simulation With Latin Supercube Sampling. *IEEE Transactions on Power Systems*, 28(2):1550–1559, May 2013.

[198] G.E. Constante-Flores and M.S. Illindala. Data-Driven Probabilistic Power Flow Analysis for a Distribution System With Renewable Energy Sources

Using Monte Carlo Simulation. *IEEE Transactions on Industry Applications*, 55 (1):174–181, January 2019.

[199] G. Chaspierre, G. Denis, P. Panciatici, and T.V. Cutsem. Dynamic equivalent of an active distribution network taking into account model uncertainties. In *2019 IEEE Milan PowerTech*, pages 1–6, June 2019.

[200] H. Verdejo, A. Awerkin, W. Kliemann, and C. Becker. Modelling uncertainties in electrical power systems with stochastic differential equations. *International Journal of Electrical Power & Energy Systems*, 113:322–332, December 2019.

[201] Chun-Lien Su. Probabilistic load-flow computation using point estimate method. *IEEE Transactions on Power Systems*, 20(4):1843–1851, November 2005.

[202] M. Aien, M. Fotuhi-Firuzabad, and M. Rashidinejad. Probabilistic Optimal Power Flow in Correlated Hybrid Wind–Photovoltaic Power Systems. *IEEE Transactions on Smart Grid*, 5(1):130–138, January 2014.

[203] C. Chen, W. Wu, B. Zhang, and C. Singh. A new point estimate method for probabilistic load flow with correlated variables including wind farms. In *2014 IEEE PES General Meeting | Conference Exposition*, pages 1–5, July 2014.

[204] M. Aien, M. Fotuhi-Firuzabad, and F. Aminifar. Probabilistic Load Flow in Correlated Uncertain Environment Using Unscented Transformation. *IEEE Transactions on Power Systems*, 27(4):2233–2241, November 2012.

[205] J. Tang, F. Ni, F. Ponci, and A. Monti. Dimension-Adaptive Sparse Grid Interpolation for Uncertainty Quantification in Modern Power Systems: Probabilistic Power Flow. *IEEE Transactions on Power Systems*, 31(2):907–919, March 2016.

[206] M. Lindner and R. Witzmann. Common Rank Approximation - A new method to speed up probabilistic calculations in distribution grid planning. In *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T D)*, pages 1–5, May 2016.

[207] S. Ackermann, M. Buhl, M. Lindner, and F. Steinke. A Novel Algorithm for the Fast and Accurate Quantile Computation in Probabilistic Power Flow. In *2018 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2018.

[208] Roger C. Dugan and Davis Montenegro. The Open Distribution System Simulator (OpenDSS), June 2019.

[209] E. McKenna and M. Thomson. High-resolution stochastic integrated thermal–electrical domestic demand model. *Applied Energy*, 165:445–461, March 2016.

[210] European Committee for Electrotechnical Standardization (CENELEC). Voltage characteristics of electricity supplied by public electricity networks. Standard EN 50160:2010, European Committee for Electrotechnical Standardization (CENELEC), Brussels, Belgium, 2010.

[211] European Committee for Electrotechnical Standardization (CENELEC). Guide for the application of the European Standard EN 50160. Standard CLC/TR 50422:2013, European Committee for Electrotechnical Standardization (CENELEC), Brussels, Belgium, 2013.

[212] Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE). Power Generating Plants in the Low Voltage Grid. Standard VDE-AR-N 4105, Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE), Germany, April 2019.

[213] D. Deka, S. Backhaus, and M. Chertkov. Structure Learning and Statistical Estimation in Distribution Networks - Part I. *arXiv:1501.04131 [cs, math]*, January 2015. arXiv: 1501.04131.

[214] S. Bolognani and F. Dörfler. Fast power system analysis via implicit linearization of the power flow manifold. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 402–409, September 2015.

[215] S. Conti, S. Raiti, and G. Vagliasindi. Voltage sensitivity analysis in radial MV distribution networks using constant current models. In *2010 IEEE International Symposium on Industrial Electronics*, pages 2548–2554, July 2010.

[216] S. Munikoti, K. Jhala, K. Lai, and B. Natarajan. Analytical Voltage Sensitivity Analysis for Unbalanced Power Distribution System. In *2020 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2020.

[217] C. Mugnier, K. Christakou, J. Jaton, M.D. Vivo, M. Carpita, and M. Paolone. Model-less/measurement-based computation of voltage sensitivities in unbalanced electrical distribution networks. In *2016 Power Systems Computation Conference (PSCC)*, pages 1–7, June 2016.

[218] P. Li, H. Su, C. Wang, Z. Liu, and J. Wu. PMU-Based Estimation of Voltage-to-Power Sensitivity for Distribution Networks Considering the Sparsity of Jacobian Matrix. *IEEE Access*, 6:31307–31316, 2018.

[219] M. Bozorg, O. Alizader-Mousavi, S. Wasterlain, and M. Carpita. Modelless/Measurement-based Computation of Voltage Sensitivities in Unbalanced Electrical Distribution Networks: Experimental Validation. In *2019 21st European Conference on Power Electronics and Applications (EPE '19 ECCE Europe)*, pages 1–9, September 2019.

[220] G. Cavraro, R. Arghandeh, G. Barchi, and A.v. Meier. Distribution network topology detection with time-series measurements. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, February 2015.

[221] M. Xu, R. Li, and F. Li. Phase Identification With Incomplete Data. *IEEE Transactions on Smart Grid*, 9(4):2777–2785, July 2018.

[222] J.D. Watson, J. Welch, and N.R. Watson. Use of smart-meter data to determine distribution system topology. *The Journal of Engineering*, 2016(5):94–101, 2016.

[223] B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In G. Dellino and C. Meloni, editors, *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, pages 101–122. Springer US, Boston, MA, April 2014.

[224] Z. Gniazdowski. Geometric interpretation of a correlation. *Zeszyty Naukowe WWSI*, 7(9):27–35, September 2013.

[225] X. Zhang and S. Grijalva. A Data-Driven Approach for Detection and Estimation of Residential PV Installations. *IEEE Transactions on Smart Grid*, 7(5): 2477–2485, September 2016. Conference Name: IEEE Transactions on Smart Grid.

[226] K. Jhala, B. Natarajan, and A. Pahwa. Probabilistic Voltage Sensitivity Analysis (PVSA)—A Novel Approach to Quantify Impact of Active Consumers. *IEEE Transactions on Power Systems*, 33(3):2518–2527, May 2018.

[227] K. Jhala, V. Krishnan, B. Natarajan, and Y. Zhang. Data-Driven Preemptive Voltage Monitoring and Control Using Probabilistic Voltage Sensitivities. In *2019 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2019. ISSN: 1944-9933.

[228] K. Jhala, B. Natarajan, and A. Pahwa. The Dominant Influencer of Voltage Fluctuation (DIVF) for Power Distribution System. *IEEE Transactions on Power Systems*, 34(6):4847–4856, November 2019.

[229] E. Vanet, S. Touré, N. Kechagia, R. Caire, and N. HadjSaid. Sensitivity analysis of local flexibilities for voltage regulation in unbalanced LV distribution system. In *2015 IEEE Eindhoven PowerTech*, pages 1–6, June 2015.

[230] T. Zhu, T. Xia, Z. Wan, and C. Zhao. A Sensitivity Analysis Method for Unbalanced Distribution Network based on Linearized Power Flow Model. In *2018 International Conference on Power System Technology (POWERCON)*, pages 1558–1563, November 2018.

[231] M. Abujubbeh, S. Munikoti, and B. Natarajan. Probabilistic Voltage Sensitivity based Preemptive Voltage Monitoring in Unbalanced Distribution Networks. *arXiv:2008.10814 [cs, eess]*, August 2020.

[232] S. Munikoti, B. Natarajan, K. Jhala, and K. Lai. Probabilistic Voltage Sensitivity Analysis (PVSA) to Quantify Impact of High PV Penetration on Unbalanced Distribution System. *arXiv:2009.05734 [cs, eess]*, September 2020.

[233] S. Munikoti, M. Abujubbeh, K. Jhala, and B. Natarajan. Spatio-Temporal Probabilistic Voltage Sensitivity Analysis (ST-PVSA)-A Novel Framework for Hosting Capacity Analysis. *arXiv:2009.08490 [cs, eess]*, September 2020.

[234] R. Dobbe, O. Sondermeijer, D. Fridovich-Keil, D. Arnold, D. Callaway, and C. Tomlin. Data-Driven Decentralized Optimal Power Flow. *arXiv:1806.06790 [cs, math, stat]*, June 2018.

[235] G.J. Fang and H. Bao. A Calculation Method of Electric Distance and Subarea Division Application Based on Transmission Impedance. *IOP Conference Series: Earth and Environmental Science*, 104:012006, December 2017.

[236] P. Lagonotte. Probabilistic approach of voltage control based on structural aspect of power systems. In *1991 Third International Conference on Probabilistic Methods Applied to Electric Power Systems*, pages 208–213, July 1991.

[237] J.A. Rice. *Mathematical statistics and data analysis*. Thomson/Brooks/Cole, Belmont, CA, 2007.

[238] M. Koivisto, M. Degefa, M. Ali, J. Ekström, J. Millar, and M. Lehtonen. Statistical modeling of aggregated electricity consumption and distributed wind generation in distribution systems using AMR data. *Electric Power Systems Research*, 129:217–226, December 2015.

[239] T.e. Huang, Q. Guo, H. Sun, C.W. Tan, and T. Hu. A deep spatial-temporal data-driven approach considering microclimates for power system security assessment. *Applied Energy*, 237:36–48, March 2019.

[240] J. Zhang, L. Chen, and P. Qin. Modeling non-stationary stochastic systems with generalized time series models. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 1061–1067, August 2015.

[241] N. Rab, F. Leimgruber, and T. Esterl. Synthetic wind speed time series with Markov and ARMA models: Comparison for different use cases. In *2015 12th International Conference on the European Energy Market (EEM)*, pages 1–5, May 2015.

[242] R. Garcia, J. Contreras, M. van Akkeren, and J. Garcia. A GARCH forecasting model to predict day-ahead electricity prices. *IEEE Transactions on Power Systems*, 20(2):867–874, May 2005.

[243] A. Cifter. Forecasting electricity price volatility with the Markov-switching GARCH model: Evidence from the Nordic electric power market. *Electric Power Systems Research*, 102:61–67, September 2013.

[244] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 1991.

[245] M. Hassanzadeh and C.Y. Evrenosoğlu. Power system state forecasting using regression analysis. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–6, July 2012.

[246] R.J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition edition, May 2021.

[247] Y. Chakhchoukh, P. Panciatici, and L. Mili. Electric Load Forecasting Based on Statistical Robust Methods. *IEEE Transactions on Power Systems*, 26(3): 982–991, August 2011.

[248] Z. Wang, M.H. Athari, and S. Hamid Elyas. Statistically Analyzing Power System Network. In *2018 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, August 2018.

[249] S.H. Elyas and Z. Wang. Statistical analysis of transmission line capacities in electric power grids. In *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, September 2016.

[250] L. Wu, S. You, J. Dong, Y. Liu, and T. Bilke. Multiple Linear Regression Based Disturbance Magnitude Estimations for Bulk Power Systems. In *2018 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, August 2018.

[251] Y. Liu, N. Zhang, Y. Wang, J. Yang, and C. Kang. Data-Driven Power Flow Linearization: A Regression Approach. *IEEE Transactions on Smart Grid*, 10 (3):2569–2580, May 2019.

[252] S.M. Mazhari, N. Safari, C.Y. Chung, and I. Kamwa. A Quantile Regression-Based Approach for Online Probabilistic Prediction of Unstable Groups of Coherent Generators in Power Systems. *IEEE Transactions on Power Systems*, 34(3):2240–2250, May 2019.

[253] M.H. Athari and Z. Wang. Statistically Characterizing the Electrical Parameters of the Grid Transformers and Transmission Lines. *arXiv:1706.02754 [physics, stat]*, June 2017.

[254] S.A. Soliman, M.H. Abdel Rahman, and M.E. El-Hawary. Application of fuzzy linear regression algorithm to power system voltage measurements. *Electric Power Systems Research*, 42(3):195–200, September 1997.

[255] A.F. Bastos, S. Santoso, V. Krishnan, and Y. Zhang. Machine Learning-Based Prediction of Distribution Network Voltage and Sensor Allocation. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, August 2020.

[256] M. Mokhtar, V. Robu, D. Flynn, C. Higgins, J. Whyte, C. Loughran, and F. Fulton. Prediction of voltage distribution using deep learning and identified key smart meter locations. *Energy and AI*, 6:100103, December 2021.

[257] R. Hadidi and B. Jeyasurya. Reinforcement Learning Based Real-Time Wide-Area Stabilizing Control Agents to Enhance Power System Stability. *IEEE Transactions on Smart Grid*, 4(1):489–497, March 2013.

[258] Q. Wang, F. Li, Y. Tang, and Y. Xu. Integrating Model-Driven and Data-Driven Methods for Power System Frequency Stability Assessment and Control. *IEEE Transactions on Power Systems*, 34(6):4557–4568, November 2019.

[259] F. Bu, K. Dehghanpour, Z. Wang, and Y. Yuan. A Data-Driven Framework for Assessing Cold Load Pick-Up Demand in Service Restoration. *IEEE Transactions on Power Systems*, 34(6):4739–4750, November 2019. Conference Name: IEEE Transactions on Power Systems.

[260] W. Greene. *Econometric Analysis*. Pearson, New York, 8th, global edition edition, 2020.

[261] B. Eisenhower, T. Maile, M. Fischer, and I. Mezic. Decomposing building system data for model validation and analysis using the Koopman operator. *SimBuild 2010*, January 2010.

[262] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, and G.M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 5th edition edition, August 2015.

[263] E. Kaiser, J.N. Kutz, and S.L. Brunton. Data-driven discovery of Koopman eigenfunctions for control. *arXiv:1707.01146 [math.OC]*, February 2021.

[264] J.L. Proctor, S.L. Brunton, and J.N. Kutz. Dynamic mode decomposition with control. *arXiv:1409.6358 [math.OC]*, September 2014.

[265] A. Banaszuk, K.B. Ariyur, M. Krstić, and C.A. Jacobson. An adaptive algorithm for control of combustion instability. *Automatica*, 40(11):1965–1972, November 2004.

[266] C.W. Rowley. Model reduction for fluids, using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos*, 15(03):997–1013, March 2005.

[267] R. Taylor, J.N. Kutz, K. Morgan, and B.A. Nelson. Dynamic mode decomposition for plasma diagnostics and validation. *The Review of Scientific Instruments*, 89(5):053501, May 2018.

[268] E. Kaiser, J.N. Kutz, and S.L. Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2219): 20180335, November 2018.

[269] T.W. Anderson and D.A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212, June 1952.

[270] NIST/SEMATECH. Anderson-Darling Test. In *e-Handbook of Statistical Methods*. National Institute of Standards and Technology (NIST), USA, 2022.

[271] T.W. Anderson. On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, September 1962.

[272] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality. *Biometrika*, 52(3-4):591–611, December 1965.

[273] R. D'Agostino and E.S. Pearson. Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1. *Biometrika*, 60(3):613–622, 1973.

[274] R.B. D'Agostino and A. Belanger. A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, 44(4):316–321, 1990.

[275] G.E. Dallal and L. Wilkinson. An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality. *The American Statistician*, 40 (4):294–296, November 1986.

[276] H. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62(318): 399–402, 1967.

[277] Y. Shi and N. Chen. Conditional Kernel Density Estimation Considering Autocorrelation for Renewable Energy Probabilistic Modeling. *IEEE Transactions on Power Systems*, 36(4):2957–2965, July 2021.

[278] S. Chatterjee and J.S. Simonoff. *Handbook of Regression Analysis*. John Wiley & Sons, May 2013.

[279] J. Durbin and G.S. Watson. Testing for Serial Correlation in Least Squares Regression. III. *Biometrika*, 58(1):1–19, 1971.

[280] G.M. Ljung and G.E.P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, August 1978.

[281] R.F. Engle. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007, 1982.

[282] M.N. Morgül Tumbaz and M. İpek. Energy Demand Forecasting: Avoiding Multi-collinearity. *Arabian Journal for Science and Engineering*, 46(2):1663–1675, February 2021.

[283] A.S. Allam, H.A. Bassioni, W. Kamel, and M. Ayoub. Estimating the standardized regression coefficients of design variables in daylighting and energy performance of buildings in the face of multicollinearity. *Solar Energy*, 211: 1184–1193, November 2020.

[284] M.A. Zamee, D. Han, and D. Won. Online Hour Ahead Load Forecasting Using Appropriate Time-Delay Neural Network based on Multiple Correlation-Multicollinearity Analysis in IoT Energy Network. *IEEE Internet of Things Journal*, pages 1–1, 2021.

[285] D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, July 1980.

[286] D.A. Belsley. A Guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4(1):33–50, February 1991.

[287] G.E.P. Box and D.R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[288] F. Petropoulos, R.J. Hyndman, and C. Bergmeir. Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, 268(2):545–554, 2018.

[289] D.P. Loucks, E.v. Beek, J.R. Stedinger, J.P. Dijkman, and M.T. Villars. *Water resources systems planning and management: an introduction to methods, models and applications*. UNESCO Digital Library, Italy, 1 edition, 2005.

[290] L. Pan. *Bootstrap Prediction Intervals for Time Series*. PhD thesis, UC San Diego, 2013.

[291] Y. Chang, J.Y. Park, and K. Song. Bootstrapping cointegrating regressions. *Journal of Econometrics*, 133(2):703–739, August 2006.

[292] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer, and M.G. Sowa. Variance reduction in estimating classification error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems*, 79(1):91–100, October 2005.

[293] B. Efron. Second Thoughts on the Bootstrap. *Statistical Science*, 18(2):135–140, May 2003. Publisher: Institute of Mathematical Statistics.

[294] J.H. Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53 (11):3735–3745, September 2009.

[295] S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer New York, NY, 1 edition, 2003.

[296] V. Cerqueira, L. Torgo, and I. Mozetic. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Mach. Learn.*, 2020.

[297] T. Dunning. The t-digest: Efficient estimates of distributions. *Software Impacts*, 7:100049, February 2021.

[298] S. Shi, X. Cheng, and P.M.J. Van den Hof. Generic identifiability of subnetworks in a linear dynamic network: The full measurement case. *Automatica*, 137:110093, March 2022.

[299] P.M.J. Van den Hof, A.G. Dankers, and H.H.M. Weerts. Identification in dynamic networks. *Computers & Chemical Engineering*, 109:23–29, January 2018.

[300] H.H.M. Weerts, P.M.J. Van den Hof, and A.G. Dankers. Identifiability of linear dynamic networks. *Automatica*, 89:247–258, March 2018.