Fault Anticipation in Distribution Networks

PhD Thesis

Eleni Tsioumpri

Advanced Electrical Systems

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

June 24, 2020

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

This thesis is concerned with the topic of Fault Anticipation in Distribution Networks, focusing on the rapidly changing operational nature of distribution networks and outlining the anticipated data-related challenges that result from these changes, encountered in practice and from related literature. With the challenge of limited data availability in mind, a data analysis methodology for Distribution Network Operators (DNOs) is presented and demonstrated through a number of short and more detailed case studies. The short case studies are illustrative examples of how the proposed methodology would be used by a DNO. In these, the identification of solar PV operation, phase imbalance and the detection of unusual network operation using dimensionality reduction are examined. The more detailed case studies form the main part of the thesis and focus on the following two areas: (i) Prediction of weather-related faults on minimally observed distribution networks and (ii) Impact of substation loading on the occurrence of power quality disturbances. More specifically, on the topic of weather-related fault prediction, the impact of weather conditions alone on the occurrence distribution network faults is explored, with the case study looking separately into the HV level (mainly 11kV - 20kV) and LV level (0.4kV) of the distribution network. The relationship of power quality events, mainly overcurrent and voltage swell events, with the load behaviour as observed at the LV side of secondary transformers is explored in the second detailed case study.

The contribution of the work presented in this thesis is twofold. First, the fact that distribution networks are currently minimally monitored or access to operational data is restricted for various reasons is acknowledged. This thesis attempts to overcome this

challenge by exploring the potential of machine learning techniques to extract valuable information from distribution networks with minimal observation. When required, the available network information is jointly analysed with data coming from different sources that can be easily obtained, such as weather observations. As mentioned above, this research has mainly focused on two areas which form the basis for the two more detailed case studies presented in this thesis. The weather-related fault prediction case study demonstrated that DNOs can predict the occurrence of weather-related faults in their distribution networks, using only weather observations from a nearby weather station and historic fault records. The other detailed case study which addressed the impact of distribution substation loading on power quality event occurrence identified a relation between representative load profiles and the transitions between them with the occurrence of power quality events. Both research subjects were selected with a common final goal in mind, which was to utilise machine learning in order to develop a methodology towards the prediction of distribution network disturbances in the absence of extensive monitoring. For the second part of the contribution, the data challenges associated with the changing state of distribution networks are assessed and suggestions to deal with these issues are made. As a result of the work presented in this thesis, an overall data analysis methodology for DNOs is proposed. The main purpose of this methodology is to identify operational or environmental factors that are more likely to lead to the occurrence of certain types of disturbances and establish relations between these factors and fault occurrence, which can then be used to predict these events. The specific case studies presented in this thesis identify relations between environmental conditions and power system faults as well as substation loading and power quality events. However, the methodology can be applied to different operating conditions and types of faults as well. Being able to establish such relations would be beneficial for DNOs as it would lead to an increased understanding of their network and allow them to act proactively in order to prevent, or minimise the impact of impending events.

Contents

List of Figures

List of Tables

Acronyms

ANN Artificial Neural Network.

BIC Bayesian Information Criterion.

BT Bagged Trees.

C2C Capacity to Customers.

CART Classification And Regression Trees.

CEDA Centre for Environmental Data Analysis.

CFSFDP Clustering by Fast Search and Find of Density Peaks.

CIM Common Information Model.

CLNR Customer-Led Network Revolution.

CML Customer Minutes Lost.

CT Current Transformer.

DFA Distribution Fault Anticipation.

DNO Distribution Network Operator.

DSO Distribution System Operator.

EM Expectation-Maximisation.

EMS Energy Management System.

ENW Electricity North West.

EV Electric Vehicle.

FCM Fuzzy C-Means.

FN False Negative.

FP False Positive.

GB Gradient Boost.

GMM Gaussian Mixture Model.

GPC Gaussian Process Classification.

HP Heat Pump.

ICT Information and Communications Technology.

ISODATA Iterative Self-Organising Data Analysis Technique.

k-NN k-Nearest Neighbour.

LCNF Low Carbon Network Fund.

LCT Low Carbon Technologies.

LDA Linear Discriminant Analysis.

LR Logistic Regression.

LSTM Long Short Term Memory.

MLP Multi-Layer Perceptrons.

MML Minimum Message Length.

NB Naive Bayes.

NPG Northern Powergrid.

PCA Principal Component Analysis.

PMAR Pole Mounted Auto-Recloser.

PMU Phasor Measurement Unit.

PV Photovoltaics.

QDA Quadratic Discriminant Analysis.

RF Random Forest.

SPEN Scottish Power Energy Networks.

SVM Support Vector Machines.

t-SNE t-Distributed Stochastic Neighbour Embedding.

THD Total Harmonic Distortion.

TN True Negative.

TP True Positive.

TSO Transmission System Operator.

VT Voltage Transformer.

WAMS Wide Area Measurement System.

WPD Western Power Distribution.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Prof. Stephen McArthur for his guidance, support and encouragement throughout this research. I would also like to thank my second supervisor Dr Bruce Stephen for his continuous support and guidance and for always finding time. I have learnt a lot from them both and, without their constructive advice, this thesis would not have been achieved.

I would also like to thank everyone at the Future Power Networks and Smart Grids CDT, particularly my 1st cohort colleagues (Anthony, Euan, James, Kyle, Marcel, Nathalie and Tania) for the nice time we had in the first year and my office neighbours Jose and Michael for the nice discussions, mostly unrelated to work.

Moreover, I would like to thank Northern Powergrid for their support and Neil-Dunn Birch in particular, for useful conversations and for providing access to the fault and power quality data. I would also like to thank Dr Steven Blair for helpful discussions and access to data. For their financial support, I would like to gratefully acknowledge the EPSRC which funded this project (Grant EP/L015471/1).

Special thanks to my best friends from Greece: Alexandra, Athina, Olga and Themis that, despite the distance, were always close to me.

I would also like to thank my parents, Anna and Vassilis, for their love and support all these years, as well as my brother Alexis and his wife Ioanna.

Finally, and most importantly, I want to thank my husband Petros for his love and encouragement and for always believing in me, and our daughter Antoinetta, who is the best thing that ever happened to me. This thesis is dedicated to her.

Chapter 1

Introduction

Energy is a basic necessity for the development of a country and the improvement of its living standards. This has been observed throughout human history and was particularly demonstrated by the progress achieved during the years following the industrialisation of our societies. However, the excessive use of energy that followed the industrial revolution, came mostly from unsustainably polluting sources such as fossil fuels and resulted in a significant increase in the concentration of greenhouse gases, especially carbon dioxide, leading to climate change. It was mainly because of the threat of climate change and its implications that decision makers around the world realised that a change in energy policy was essential. The carbon reduction targets that many countries have committed to achieve, combined with the increasing demand in electricity, have created the need for a more sustainable way of producing and consuming electrical energy. Today's grid, which is comprised to a great extent of aged infrastructure, cannot support all the required changes and in order to develop a sustainable energy sector, the transformation of the electrical grid is essential. The next generation grid, the so called "smart grid", will be able to integrate large amounts of renewable energy and, most importantly, it will allow the power and information flow between producers and consumers, which will create new opportunities for the electricity grid. This transition to the future smart grid involves significant changes, the largest part of which will occur in the distribution network.

The electrical network in the UK, as in many other countries, is currently undergoing this transformation process which will facilitate the widespread use of Low Carbon Technologies (LCTs), while maintaining the quality and security of supply. This has led to an increasing amount of monitoring equipment being installed at the distribution level in the past few years, resulting in more data being available for processing. The different types of data generated by these monitoring devices have the potential to give an unprecedented view of the network and, with appropriate processing, can help utilities significantly improve their services. As all this is relatively new, however, Distribution Network Operators (DNOs) have still a long way to go in order to be able to fully utilise this data and reap the benefits associated with it. The increased data availability comes with a number of challenges for the DNOs, as the existence of new monitoring devices in the network does not automatically mean enhanced visibility of the network, in the sense that it heightens understanding of its operating state. Currently, it is common for DNOs to have monitoring equipment in place and never look into the recorded data. In order to extract value from the increasingly available data, a number of actions regarding the management and analysis of this data need to be taken by the DNOs.

The contribution of the work presented in this thesis is twofold. First, the fact that distribution networks are currently minimally monitored or access to operational data is restricted for various reasons is acknowledged. This thesis attempts to overcome this challenge by exploring the potential of machine learning techniques to extract valuable information from distribution networks with minimal observation. It is worth clarifying here, that the phrases “minimally monitored” and “minimally observed” that are used throughout this thesis refer to networks with either no monitoring in place or with limited data availability and should not be confused with their meaning when they are used in the context of other power system areas, such as state estimation. When required, the available network information is jointly analysed with data coming from different sources that can be easily obtained, such as weather observations. Machine learning has been selected as a tool for the proposed data analysis methodology for various reasons. First, the intention is that the proposed methodology can be applicable now and

in the future, where very large amounts of data are expected to be available for a distribution network. Using machine learning, the task of identifying trends and patterns within large volumes of data has become much easier. In addition, one of the main advantages of machine learning is that the models can independently adapt when they are exposed to new data. The ability of these models to improve over time is very important, as the developed models can continuously improve their predictions by taking into account new observations. This research has mainly focused on the following two areas: (i) weather related fault prediction in minimally observed distribution networks and (ii) impact of distribution substation loading on power quality event occurrence. Both research subjects were selected with a common final goal in mind, which was to utilise machine learning in order to develop a methodology towards the prediction of distribution network disturbances in the absence of extensive monitoring. For the second part of the contribution, the data challenges associated with the changing state of distribution networks are assessed and suggestions to deal with these issues are made. As a result of the work presented in this thesis, an overall data analysis methodology for DNOs is proposed. The main purpose of this methodology is to identify relations between network or environmental factors and power system faults or power quality disturbances and eventually predict these events. Being able to establish such relations would be beneficial for DNOs as it would lead to an increased understanding of their network and allow them to act proactively in order to prevent, or minimise the impact of impending events.

1.1 Principal Contributions

The novelty and contributions of this research can be summarised in three main points. First, a general data analysis strategy for DNOs is proposed, where statistical analysis and machine learning algorithms can be used to repurpose existing data and extract additional value through case studies stemming from tacit domain knowledge. The "general" data analysis strategy and the "overall" data analysis methodology differ in that the "general" strategy involves both the data management at the DNO level and the subsequent data analysis, while the "overall" methodology refers to a data analysis

methodology, which covers the different stages of data analysis and is applicable to the different kinds of data available for a distribution network. Secondly, using this strategy, this thesis demonstrates that DNOs can predict the occurrence of weather-related faults in their distribution networks, using only weather observations from a nearby weather station and historic fault records. Finally, a relation between representative load profiles and the transitions between them with the occurrence of power quality events is identified in this thesis. The three main points of contribution discussed above and their impact are presented below:

- A review and assessment of the data related challenges associated with the transformation of the electrical grid is presented. Using the information derived from literature and the lessons learned during this PhD project, the main challenges are identified and recommendations for a more efficient data management strategy are made. With the identified challenges in mind, a general data analysis methodology for DNOs is proposed and demonstrated through a number of short and more detailed case studies, providing examples of how data analysis methods can utilise existing data to increase visibility and improve network operation.

- A methodology for finding the functional relation between distribution network fault occurrence and environmental conditions is demonstrated. It is shown that it is possible to predict the occurrence of a weather related fault, considering only weather variables rather than detailed topographic and meteorological information. This could be used by DNOs to improve the visibility of their unmonitored networks as they could use longer term weather forecasts to identify vulnerable parts of the network under the forecasted conditions and plan their maintenance strategy accordingly.

- The recurrent load profiles of distribution substations and the transitions between them are used to predict the occurrence of power quality events, mainly voltage swells. This is important as DNOs could identify potential constraint violations in advance, which could inform them how much capacity can still be connected on the network, something that is aligned with their future business objectives.

1.2 Thesis Outline

The remainder of this thesis is organised as follows:

In Chapter 2, the current state of distribution network monitoring is presented and an overview of the previously conducted research in the wider area of fault prediction and event detection is given. Then, recent advances in academic research relevant to the more specific topics of weather related fault prediction and load behaviour generalisation are discussed.

Chapter 3 discusses the challenges arising from the rapidly changing distribution network as seen in the literature as well as encountered throughout this project. Recommendations on what actions should be taken by the DNOs in order to cope with the anticipated increase of distribution network data are also made.

In Chapter 4, the proposed Data Analysis Methodology is described stage by stage. The data analytics methods used throughout this thesis are presented and explained along with the assessment criteria considered.

In Chapter 5, a set of short case studies involving the application of machine learning methods to distribution network data in order to identify signs of unusual operation are described. The purpose of these case studies is to provide brief examples demonstrating how the proposed methodology could be used by a DNO.

Chapter 6 addresses the topic of weather related fault prediction in minimally observed distribution networks using only weather observations. Using historical fault records from a UK DNO's licence area, the ability of various classification methods to find a relationship between weather conditions and fault occurrence was investigated. The data analysis process that was followed is detailed in this chapter, where the results obtained from the classification methods are compared and the best performing methods with respect to fault prediction are identified. The potential applications in the distribution network and the limitations of the derived predictive models are also discussed in this chapter.

Chapter 7 explores the impact of distribution substation loading on the occurrence of power quality disturbances. The topic of power quality event prediction given the substation load behaviour is addressed by clustering the daily load profiles from distribution substations in order to derive the representative load profiles for these substations. Each day is then assigned to the representative profile that is more similar to the observed profile. The identified load profiles and the transitions between them are used as inputs to a machine learning model in order to predict the occurrence of power quality events on a particular day. The methodology presented in this chapter is demonstrated with two approaches. First, each substation data is clustered separately and the representative load profiles are obtained per substation. In the second approach, the load profiles from all available substations are used to obtain the representative profiles across substations. The performance of power quality event prediction, which is performed per substation in both cases, is compared and discussed in this chapter.

Finally, in Chapter 8, a brief summary of the contents of this thesis is provided and concluding remarks with focus on the main contributions of the research are presented. Possible areas where the relevant future work could focus are also discussed in this chapter.

1.3 Publications

E. Tsioumpri, B. Stephen, N. Dunn-Birch and S. McArthur, "Data Analytics to Support Operational Distribution Network Monitoring", In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*

E. Tsioumpri, B. Stephen and S. McArthur, "Weather Related Fault Prediction in Minimally Observed Distribution Networks", submitted for review

E. Tsioumpri, B. Stephen and S. McArthur, "Substation Duty Cycle Impact on Distribution Fault Occurrence", submitted for review

Chapter 2

Distribution Network Monitoring

Traditionally, the electricity network had a centralised structure, where the energy was produced in large central power plants and was then transmitted to load centres and distributed to the consumers. For many years, the distribution networks have been a passive part of the electricity grid since their main role was to accommodate the uni-directional power flow from substations to consumers. This is now changing, as new technologies are increasingly being installed in the distribution network in order to meet the carbon reduction targets, as discussed in the Introduction. The large scale adoption of LCTs such as solar Photovoltaics (PV) and wind generation, Electric Vehicles (EVs), Heat Pumps (HPs) and other technologies will significantly increase the distribution system's complexity and raise a number of issues regarding the operation and management of distribution networks. In order to integrate these changes efficiently and address the new challenges without compromising the quality and security of supply, the future distribution networks are expected to heavily rely on monitoring systems and data utilisation to develop analytics-based solutions to assist in network operation.

The transformation of the electrical grid is illustrated in Figure 2.1, which shows the traditional grid structure, where the power flow is unidirectional (top) and the struc-

ture of the future grid, where bi-directional power and information flow is expected (bottom).

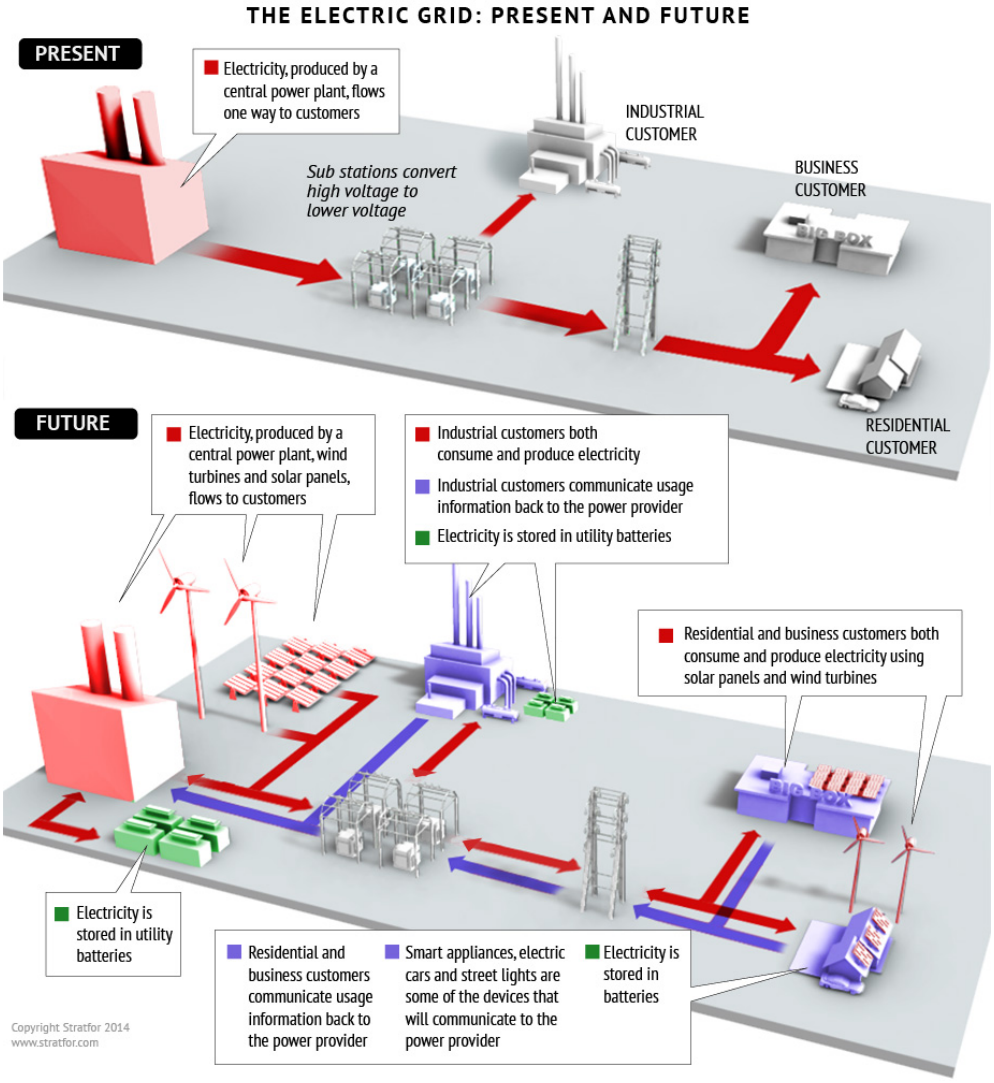


Figure 2.1: Traditional and future electricity grid [1].

The purple arrows in the bottom part of the figure represent usage information flow between the different components of the future grid. These data sources combined with the additional equipment that is expected to be installed in the distribution network in order to monitor the condition of network components and to ensure the normal operation of the network will produce vast amounts of data. Thus, it is essential for DNOs to be able to use the valuable information contained in this data to support the

operation of their network. The increased information flow at the distribution level will also provide the future Distribution System Operators (DSOs) with service capability and will facilitate the cooperation between DSOs and Transmission System Operators (TSOs) in order to balance the power system as a whole. According to the Energy Networks Association, the Distribution System Operator is defined as follows [2]:

"A Distribution System Operator (DSO) securely operates and develops an active distribution system comprising networks, demand, generation and other flexible distributed energy resources (DER). As a neutral facilitator of an open and accessible market it will enable competitive access to markets and the optimal use of DER on distribution networks to deliver security, sustainability and affordability in the support of whole system optimisation. A DSO enables customers to be both producers and consumers; enabling customer access to networks and markets, customer choice and great customer service."

As DNOs are currently in the transition phase towards the future DSO, this definition will evolve as the industry develops. The increasingly installed generation at the distribution level means additional responsibilities for the DSO compared to those of a DNO. In particular, ancillary services that have been traditionally provided by large generating units connected to the transmission system, will need to be provided by generators connected in the distribution network as well. To select the suitable congestion management measures and achieve system balancing in a cost effective manner, a strong cooperation between all system operators will be necessary [3]. The increased responsibilities of the DSO as well as the associated increase in the levels of cooperation discussed above are pointed out in the ENA DSO Transition Roadmap [4], in which data utilisation and analytics have a prominent role. Therefore, being able to effectively manage the anticipated data flows at the distribution network is crucial to maintain a reliable and secure grid. The developments outlined above have created new opportunities for scientific research in the field of power systems. As the main topic of interest for the work presented in this thesis is the prediction of faults or other unwanted events in distribution networks, providing a detailed presentation of the re-

search that involves distribution network monitoring and data utilisation is beyond the scope of this thesis. In the next sections, the previously conducted research addressing the topics of (i) fault prediction and power quality event detection in distribution networks, (ii) weather-related fault prediction and (iii) generalisation of load behaviour is presented. As the research undertaken and presented in this thesis focuses mainly on the prediction of weather related faults and on the impact of substation loading on the occurrence of power network disturbances, this chapter will mainly focus on the relevant areas. A comparison with the existing literature, which highlights the main points that distinguish the work presented in this thesis from the previously conducted research, is given in the last section of this chapter.

2.1 Fault Prediction and Power Quality Event Detection in Distribution Networks

Fault prediction in distribution networks is a broad area of research and a number of works, focusing on different aspects of this topic, have been recently published. This section will provide a brief presentation of the relevant research works, focusing on the data analytics and machine learning methods that have been used.

Selected examples of research work on data analysis of electrical network monitoring data are presented in this section, which contains research works on fault prediction as well as the use of data analytics for distribution network applications in general, giving an overview of the areas currently covered by the relevant research.

Researchers in the Power System Automation Laboratory in Texas A&M University, led by B. Don Russell, have been working on fault anticipation since the late 1990's. The aim of their Distribution Fault Anticipation (DFA) project was to increase network reliability by anticipating events that could cause outages and provide utilities with a valuable tool that would allow their timely intervention and prevention of catastrophic failures [5]. The project involved the development of prototype monitoring devices as well as the necessary algorithms to process the collected data. A number of utilities participated in the project and the DFA monitoring systems were installed on their sub-

stations on a per feeder basis, monitoring voltage and current waveforms as measured by conventional Current Transformer (CTs) and Voltage Transformer (VTs) on the feeders. The algorithms developed are generally based on advanced signal processing and pattern recognition. They are used for recurrent fault clustering and generate alerts, when the measured current or voltage measurements match those that have previously been identified as faults. However, the analysis does not require specific methods and a number of different techniques have been used, such as examination of RMS voltage or current levels, temporal or spectral analysis of voltage and current waveforms, fuzzy logic, Bayesian network etc [6]. The waveform analytics functions of their fault anticipation system are performed entirely at the device level in the distribution substations and a number of real-life applications can be found in [7], [8].

Artificial Neural Networks (ANN) have been considered for fault prediction in transmission lines since 1990's, as it can be seen in [9]. The aim of the technique developed in this work was to identify a fault at early stages and notify the operator in order to take the necessary actions to prevent it. A neural network based detector was developed to monitor equipment such as switchgear and transformers, and used external measurements from input and output nodes of the system. The training data set for the ANN were generated by simulations and included a variety of input fault states, even small changes that would normally not be sufficient to trigger protection. It was shown that early detection of faults could be achieved by analysing these small changes before they evolve to major events.

A very recent application of neural networks in a fault prediction system is discussed in [10], which addresses the topic of line trip fault prediction in order to prevent large scale outages that usually result from this type of faults. In this work, feature extraction is performed using Long Short Term Memory (LSTM) networks, which are a type of recurrent neural networks, while Support Vector Machines (SVM) are used for the actual prediction of line trip faults.

The potential of automatically analysing Pole Mounted Auto-Recloser (PMAR) data in order to predict faults and reduce the number of customer interruptions has been

studied in [11]. The aim of this work, which combines expert knowledge with data mining in order to produce diagnostic and predictive rules, was to identify pre-fault activity and semi-permanent faults causing frequent intermittent outages, which are known as nuisance tripping, and prevent potential permanent faults caused by the same reason. The proposed system uses SCADA alarms and data from NOJA PMARs installed on overhead lines in the distribution network of the Scottish Power Energy Networks (SPEN). When an event occurs, the PMAR records relevant information such as open-close sequence, types of faults detected and currents on affected phases. In addition, data from other sources are utilised to support the analytical process which is briefly discussed below. First, the SCADA alarms are examined to identify if there was any nuisance tripping. Then the proposed solution identifies if there was a semi-permanent fault that caused a nuisance tripping and extracts the related pre-fault activity. After the patterns and trends of the identified semi-permanent faults are detected, historical data are utilised in order to cluster the identified semi-permanent faults and their signatures and predict potential permanent faults.

Condition-based maintenance of different assets, through monitoring of the assets and data analysis, also involves some type of prediction, as it can detect incipient failures, and prevent them from evolving to major events. A decision support system analysing trip coil current signatures obtained from off-line distribution circuit breaker tests is proposed in [12], where data mining is combined with expert knowledge to provide assessment of the circuit breaker condition given a trip coil current signature. The normal operation of a circuit breaker produces a specific trip coil current signature. The data-driven approach proposed in this paper, analyses field data captured by operating circuit breakers and identifies various trip signatures. The observed trip coil signatures form clusters which represent different circuit breaker conditions. The proposed system uses expert interpretation of the observed trip signatures in order to produce simple output signals to inform the field technicians of the normal, defective or faulty condition of the circuit breaker. This gives the utility the opportunity to intervene immediately, in case of a faulty condition, or in the case of a defective condition, before the next scheduled maintenance and prevent failure. Similar work has been done

in [13] where an automated analysis of on-line trip coil current data is proposed. This work utilises signal processing and expert systems in order to identify anomalies in the recorded circuit breaker operation. An example of vibration pattern analysis is given in [14], where continuous vibration (but not on-line) data associated with the circuit breaker operations are analysed in order to identify signs of faulty conditions within the data.

Another example of fault anticipation related research can be found in [15], where machine learning susceptibility analysis was used to predict failures on distribution feeders. In this work, which was applied to approximately 1000 distribution feeders in the area of New York, a machine learning system was developed to rank the feeders based on their susceptibility to failure. A new machine learning algorithm was developed for this application which focused on feeder failures due to overloads. This system aimed to accurately rank the most vulnerable feeders, which would allow the system operators perform preventive maintenance in order to maintain the systems reliability under increasing load conditions. The system proved to be successful as the feeders ranked as the most susceptible were responsible for approximately 40% of the failures.

Electrical network data has been analysed for a number of different purposes apart from fault prediction, such as examining the impacts of different technologies on the network, detecting and classifying different events or to identifying representative networks in order to accelerate and simplify the analysis. Examples of research related to this type of analysis of electrical network data are given below.

In 2009, Asheibi et al [16] analysed data collected by a harmonic monitoring program previously implemented in a distribution network in Australia. The purpose of the analysis was to investigate the impacts of harmonic distortion problems on the network caused by the increased use of power electronic connected equipment at the distribution level. The presence of excessive harmonics can cause power quality issues, increased current in the neutral conductor as well as losses and can even lead to equipment failure. The data used for the analysis covers a duration from August 1999 to December 2002 and consists of fundamental, 5th and 7th harmonic orders of voltage and cur-

rent, as well as total harmonic distortion for each of the three phases. Data mining tools and techniques were utilised in order to identify clusters in the data that represent different operating conditions that could be used to detect anomalies. Supervised learning makes use of labelled data to learn a relation between known input and output variables, while the goal of unsupervised learning is to identify underlying patterns within the data. Both supervised and unsupervised learning are used in this work. Unsupervised learning is used to identify natural clusters in the data. To find these clusters, which represent different operating conditions, a method based on Minimum Message Length (MML) is proposed. Subsequently, the C5.0 algorithm is used for the supervised learning, which generates a decision tree or a rule structure to describe the clusters using the data attributes. Their approach was able to recognise a number of clusters representing different operating conditions or events such as: overloading conditions, capacitor switching events, on-peak and off-peak periods, peak use of A/C and turning on of off-peak water heaters.

In 2013, researchers from Eindhoven University and Alliander energy network company in Netherlands used an approach based on the work developed by [16] in order to utilise the large amounts of data collected by Alliander's distribution network and investigate the potential of data mining applications in power quality data evaluation [17]. The data analysed in this study were not from a real network but from simulations representing a typical distribution network in the Netherlands. Fundamental, 5th and 7th harmonic currents at 12 sites on the network were considered in the analysis, which utilised additional data such as reactive power to confirm the suspected events. A number of different operating conditions were identified during this analysis, with three of them representing normal operating conditions and four representing transient moments. The source of a harmonic distortion, the route of major harmonic current owing, capacitor switching events and location of compensation capacitors were identified by their approach.

In order to investigate the impacts of low carbon technologies, such as electric vehicles and photovoltaics on the distribution network, Rigoni et al [18] used clustering tech-

niques to obtain representative feeders and accelerate the analysis. Different types of data including current, power, customer details, presence of residential scale generation etc, were considered for 232 LV feeders from the North West of England. Four clustering algorithms were used and compared using a number of validity indices in order to obtain the representative feeders. A total of 11 representative feeders were obtained (8 without PV generation and 3 with PV), which were then used to assess the PV hosting capabilities of all the feeders. It was shown that using the representative feeders, it was possible to identify the level of PV penetration at which specific types of feeders are more likely to present problems.

Lazzaretti et al [19] proposed a method for identification and classification of events using voltage waveform data. The method was also able to detect events that were not seen before, mentioned as novelties. For this study, simulation data from 621 simulated events were analysed and 29 classes of events were considered. A number of one-class classifiers were used and compared during this study and were applied to this multi-class problem using two approaches: (i) Independent multi-class classification and novelty detection, which involved separate stages for classification and novelty detection, and (ii) Coupled multi-class classification and novelty detection, where the two happened at the same time. It was found that the second approach performed better in most cases. In 2016, Lazzaretti et al [20] extended their previous work using both simulated and real voltage waveforms. Using the same classifiers as those used in [19] but in a modified version of their previously developed methodology that allowed the model to learn from a small number of examples for each class, they achieved increased classification and novelty detection performance.

Another method for the classification of multiple events in real time was proposed in [21]. Their method was moving window Principal Component Analysis (PCA) and the moving window was used to provide different thresholds for event detection, which allows the model to identify the normal variations of the power system behaviour and adapt to the different normal operating conditions throughout the day. Real and simulated data from Phasor Measurement Units (PMUs) were used for this analysis. PMUs

are advanced monitoring devices that can measure the magnitude and phase of voltage and current phasors that enable the comparison of data obtained from distant devices due to the time synchronised measurements they provide. The model developed in this paper was able to distinguish between islanding and generation-load mismatch events, such as line trips, interconnection trips, generation dip and loss of load. However, it could not distinguish between multiple concurrent loss and generation events, i.e. when several events (related to loss of either load or generation) occurred at the same time, the proposed method could only classify the resulting over- or under-generation.

A more general approach for the utilisation of PMU data and the detection of disturbances was proposed in [22], where the concept of Wide Area Measurement System (WAMS) was extended from transmission to distribution level for the first time. The aim of their work was to utilise data from devices with PMU functionalities to improve situational awareness. A number of data visualisation and analytics tools and techniques were implemented to process the large volume of data, through real time applications as well as offline data analysis. The real time applications involved visualisation of the measurements, disturbance detection and identification of location, islanding and off-grid detection and model updating based on real time measurements among others, while the non real time applications involved functions such as post event analysis, statistical analysis of historic data, model validation etc.

As fault prediction is a very broad and active field, selected research examples which were the most relevant to the topic of this thesis, were presented above. The following two sections provide details of the previously conducted research in the areas of weather related fault prediction and load behaviour generalisation as these are the main areas explored in this thesis.

2.2 Weather-Related Fault Prediction

Adverse weather conditions can have a significant impact on the electricity network infrastructure and, subsequently, compromise the quality of power delivered to consumers. A study on the effects of climate change on the US electrical network concluded that

80% of all large scale power outages between $2003 - 2012$ were caused by weather and the average number of weather related outages per year doubled during those years [23]. Although some results refer to weather conditions specific to the US climate, they are indicative of how the changing weather conditions can affect the electricity network. In the UK, the distribution network operators have published climate adaptation reports, outlining the current risks and the anticipated impacts as a result of a changing climate. Among others, [24] discusses the main results of a study conducted with the UK Met Office regarding the impacts on the electricity network, which identified the major causes of weather related outages and estimated how their frequency might change in the future. Using the Met Office climate projections [25], the study showed that there is an uncertainty regarding the future occurrence of wind related faults, as there is uncertainty in the wind gust projections as well. However, the number of lightning related faults is more likely to increase and the faults due to snow, sleet and blizzard are estimated to be fewer but with the same or increased intensity. A review of the research addressing the impacts of extreme weather on the power systems' resilience is presented in [26], where a framework for the modelling of weather related impacts on power systems is proposed. A methodology based on this framework is developed in [27], where the effects of windstorms on the transmission network's resilience are assessed. In order to do this, real time weather conditions are taken into account and the weather dependent failure probabilities are calculated. The application of this methodology on the GB transmission network determined the critical wind speed below which the network was very robust. However, above that wind speed there was a sharp increase in the event occurrences per year.

The effect of wind on the GB transmission system was also investigated in [28], where historical data was used to identify the relationship between wind gust and fault occurrence. The work presented in this paper concluded that, when extreme values of wind gust are observed there is a higher probability for a wind related fault to occur. The occurrence, intensity and duration of wind storms in the northeast US are modelled in [29]. Subsequently, the dependencies of weather and component failure are investigated, and the risk of failure is quantified for the components of a real distribution

system.

Weather data has been used as part of an improved protection strategy called hierarchically coordinated protection [30], [31], which can be applied to both the transmission and the distribution networks. Unlike other fault prediction approaches which aim to prevent the occurrence of a fault, the purpose of prediction in hierarchically coordinated protection is to give the utilities the opportunity to anticipate a weather-related fault and be better prepared to deal with it. This approach utilises weather data and machine learning techniques such as neural networks or support vector machines to detect and classify the potential faults. Then, when a fault is detected and recognised by the system, the protection is adjusted based on the type of the fault. The prediction of occurrence and location of weather-related faults was also examined in [32], which focuses on the distribution network and provides a comparison of machine learning models developed for this purpose. Again, the aim of these predictive models, which utilised grid electrical parameters and infrastructure type alongside historical weather and fault data, was to enhance preparedness for an event rather than preventing it.

The use of historical weather data alongside a number of other data sources such as customer calls and smart meter data, geographical information system data, asset condition data etc, for post fault analysis is proposed in [33]. Work utilising the above ideas is presented in [34] and [35], where historical and real time weather data is analysed alongside data from various other sources in order to provide an understanding of the effects that different nature-caused events have on the network and produce risk maps for weather-related outages using a geographical information system framework and fuzzy logic respectively.

Weather conditions and lightning strike positions have been used in addition to data from remote power quality monitoring devices to improve their predictive maintenance system by detecting incipient equipment failure in [36], while in [37] wind speed data in conjunction with component resiliency index and distance from the hurricane centre have been used as inputs to a support vector machine model, in order to predict an electrical grid component outage following a hurricane.

The modelling approach presented in [38], involves analysis of data coming from various sources such as maintenance tickets, information on equipment vulnerability and weather variables, mostly related to temperature and precipitation. The aim of this model was to gain an insight of the weather factors that significantly affect the power grid and, subsequently, lead to serious events and model their dependencies.

A framework to predict the duration of distribution system outages is presented in [39]. Using outage reports and their respective repair logs in conjunction with weather data, it was found that certain weather features were correlated with specific causes and good results could be achieved, even when taking only weather data into account. The inclusion of information contained in the outage reports and repair logs was found to enhance the model's performance.

An analysis of the correlation between failures and weather conditions is presented in [40], where market basket style analysis is used to generate predictive rules using weather data which has been earlier categorised as "high", "medium" or "low". The analysis gave a moderate accuracy of prediction but indicated that there is potential in using weather forecasts to predict component failure. In [41], an extended version of logistic regression is used to perform a probabilistic classification and calculate the probability of a fault occurrence given information regarding the weather conditions, location, time and operating voltage.

The relationship between weather conditions and the total number of interruptions is examined in [42], where historical weather data and daily number of failures are considered in order to predict the total number of weather related failures in a year. The purpose of this work is to assess the network's performance in the end of the year by comparing the actual and the previously predicted number of failures. Similar work is presented in [43], where a neural network based model is developed to predict the total number of interruptions and not only those that are directly caused by weather conditions.

2.3 Generalisation of Load Behaviour

Earlier, more fundamental research on the substation load behaviour showed that particular loads behaved in particular ways. This was demonstrated in Western Power Distribution's (WPD) LV Templates innovation project that was completed in 2013. The purpose of this project was to use load behaviour information from selected representative substations and develop templates that could be applied to the unmonitored substations as well [44]. Using k-means clustering they grouped the selected substations into the developed clusters based only on their load profiles. Then, they used fixed substation data, such as substation capacity, number of customers, number of connected PV generation etc to characterise these clusters. Finally, they assigned unmonitored substations to the above clusters using Multinomial Logistic Regression. Similar work, albeit not as detailed as the LV templates project, has been presented in [45], where distribution substation load data were analysed and clustered in order to obtain representative load profiles and in [46] where the load profiles derived from clustering of substation data are used alongside the energy consumption of customers in order to determine the electricity demand on each bus.

The above works apply clustering methods to group load profiles as measured at the distribution substations. However, the clustering of load data measured at individual customers' premises using smart meters has been more thoroughly investigated. The different techniques such as k-means clustering, hierarchical clustering, Gaussian Mixture Models, fuzzy clustering methods and others that have been used for load profile clustering and the various applications that stem from these works related to load forecasting, tariff formulation, bad data detection, demand response etc are discussed in [47], [48] and [49]. Among the techniques used for load profile clustering, k-means has been the preferred method as the majority of works have used the original or modified versions of it.

In [50], an adaptive k-means clustering method was applied to hourly smart meter data collected from a US network to group similar load profiles. Then, a hierarchical clustering approach was used to merge clusters with very similar characteristics. The

purpose of this work was to identify which customers had the highest potential for demand response actions. Improving load forecasting and subsequently selecting the best demand response options was the motivation for the work presented in [51], where a two-level load profiling model based on k-means clustering was proposed. In the first stage, the load profiles for the local power consumption are derived and, then, they are used in the second stage for the development of a global power consumption profile. A methodology based on adapted versions of k-means, Kohonen adaptive vector quantization, fuzzy k-means and hierarchical clustering, which can be used for load forecasting and load determination, is proposed in [52]. Another k-means based clustering model is described in [53], where the aim is to improve the system level intraday load forecasting.

Gaussian Mixture Models (GMM) are another method that has been used for load profile analysis and clustering but it is not as common as k-means clustering. In [54], a statistical representation of the load using GMM is proposed. After a comparison with a number of other statistical distributions, they proved that GMM could accurately represent the electrical load. Half-hourly smart meter data were clustered using GMM in [55], as discussed later in this section, while [56] used GMM to cluster 1300 load profiles and then applied Markov Models to generate synthetic load profiles that could be used for further analysis, addressing in this way the privacy concerns that are often associated with the smart meter data.

Other methods that have been used for load profile clustering include: Iterative Self-Organising Data Analysis Technique (ISODATA) described in [57] to achieve more accurate network calculation, the centroid method used in [46] to improve the planning and operation processes of distribution networks, a frequency domain based clustering technique proposed in [58] to reduce the number of features and speed up computations. Self organising maps [59], modified follow the leader clustering [49], Fuzzy C-Means (FCM) [60], Clustering by Fast Search and Find of Density Peaks (CFSFDP) [61] and wavelet-based clustering [62] can also be found in the relevant literature.

From the research presented above, it is evident that load profile clustering is a popular

topic among power systems research groups. However, in the majority of relevant research, static approaches to address load profiling are used, assuming that the load is consistent. What is not being frequently done is identifying the dynamics of load behaviour and use the knowledge that resides in the load dynamics to improve network operation. This aspect of load profile clustering was addressed by Stephen et al. and Wang et. al in [55] and [61] respectively. Building on the basic concept of load profile generalisation that was well established by previous works, the authors in [55], used GMM to identify recurring load profiles for 32 residential properties for one month using half-hourly smart meter data. Looking further into these profiles to understand how these behaviours change over time, they found that some loads behave in the same way all the time, while others change over time. Potential applications of their results, such as using the load variability to improve load-flow calculations were discussed. In [61], the authors used a density based clustering algorithm called CFSFDP to group the customers into clusters with similar dynamic characteristics. Subsequently, they used the clustering results in order to identify the demand response potential of the customers in each group.

2.4 Contributions of this Research in the Context of Prior Work

The day ahead prediction of weather-related faults in distribution networks with minimal monitoring and the impact of substation loading on the occurrence of power quality disturbances are the two main topics of this thesis. The relevant literature was presented and discussed earlier in this chapter. In this section, the gaps identified in the previously conducted research and the differences with the work presented in this thesis are presented and discussed.

Before going into the two specific topics that are being addressed in this thesis, it is worth noting that the practical challenges around data management in a DNO or related utility are taken into consideration. These challenges refer to those identified in the literature or encountered in practice and this thesis gives recommendations that

would facilitate the use of data analytics for applications such as fault prediction. None of the prior literature related to the topics discussed above detail these practical challenges, while much of the relevant research work in the field utilises simulated data or real data that has been gathered for specific purposes. The data challenges and the recommendations for an improved data management approach that would allow the DNO to repurpose existing data and, therefore, maximise its value are discussed in Chapter 3.

The research work presented earlier in this chapter, gives an idea of how weather data has been used to predict weather related faults for various applications. The first examples of research work discussed, refer to fault prediction at the transmission level and wind is the environmental factor that is predominantly considered. Next, the relevant work at the distribution level was discussed. The majority of this work utilises a substantial amount of data, coming from various sources alongside the weather data that they use. Two of the papers presented above [42], [43], make use of weather data and number of failures only but their purpose is to predict the total number of interruptions in a region. Another [40], aims to predict a component failure using only weather data but, instead of using the actual measurements, they have previously classified them in three categories (high, medium, low). In contrast, the research work presented in this thesis aims to assess the impact of weather on the occurrence of distribution network faults in the absence of extensive monitoring. As the distribution networks in the UK are usually minimally observed, this work utilises only already existing meteorological data from local Met Office weather stations and fault records provided by a UK Distribution Network Operator, in order to predict a weather related fault occurrence at a given location. A detailed presentation of the relevant analysis and results are presented in Chapter 6.

Regarding the generalisation of load behaviour, the works presented above show that the load behaviour follows certain patterns which can be described by a number of representative load profiles. Additionally, they show that changes in load behaviour over time have been observed. The work presented in this thesis extends the work

done previously, by trying to identify how these changes in load behaviour can be used to improve network operation. Distribution Network Operators have observed the occurrence of many faults in the distribution network during transitions from spring to summer and from summer to autumn. This thesis explores the way in which substation duty cycles can affect the network and attempts to find a relationship between the changing load profiles and the occurrence of faults or disturbances. To do that, low resolution substation data are jointly analysed with records of power quality disturbances that manifest at higher resolutions. A detailed presentation of the relevant analysis and results are presented in Chapter 7.

Chapter 3

Data Challenges and Requirements

The distribution network is the part of the electricity network that transfers electricity from the transmission grid to the end users and covers a wide range of votages. In England and Wales the distribution voltages range from 230 V to 132kV, while in Scotland, 132kV is part of the transmission network. This chapter is looking into the data related challenges at the distribution level, focusing mainly on the voltage levels at 33kV and below. These networks are much less monitored than the transmission level, with the lower voltages at the distribution being usually minimally observed, meaning that little to no data regarding the network operation is available for large parts of these networks. These unmonitored parts of the network are usually in the last mile, where a significant amount of LCTs is connected. In addition, in areas of the distribution network where monitoring is available, the collected measurements are either used for specific purposes (for example to investigate the impact of certain technologies on the network, to determine what is happening at a specific area of interest, to identify a location of a fault after it has occurred etc) or are not used at all. This was also the case in the last few years, when large amounts of monitoring equipment was installed in parts of the distribution network as part of various smart grid demonstration projects undertaken by the UK DNOs. During these projects, only part of the monitoring data

was analysed and for specific purposes, while some of the monitoring devices continued collecting data, which was never extracted and used for further analysis, even after the completion of these projects. The lack of a DNO strategy for utilising the data available on their networks is evident from the above. As discussed earlier in this thesis, future distribution networks are expected to have extensive monitoring meaning that in order to be able to cope with the changing nature of the distribution networks and the increased responsibilities resulting from those changes, developing a framework for data management and analysis is essential for a DNO. This thesis aims to help with this by providing a set of recommended actions regarding the management of data and a general data analysis methodology on how they can utilise their data to support network operation now and in the future. The methodology is presented in the next chapter and is demonstrated by the case studies in the following chapters. The focus of this chapter is on the data requirements and challenges as well as the recommended practices for DNOs.

3.1 New Data Streams in Future Distribution Networks

Currently, the parts of the distribution networks that are being monitored include mostly equipment at the HV level of the network such as [63]:

- Primary substation transformers (33kV/11kV, 33kV/6.6kV), where current and voltage are typically monitored using CTs and VTs respectively. Half hourly averages of these quantities as well real and reactive power are typically stored in a database to be used for planning purposes.
- HV feeders (6.6kV, 11kV), which are monitored at the primary substation and selected locations along the feeder. Current is the quantity that is typically measured at an HV feeder.
- HV connected customers, where CTs and VTs are used to monitor current and voltage at half-hourly resolution.

Unlike primary substations, the majority of secondary substations are not currently

closely monitored. This refers to the low voltage side of the secondary substations, as opposed to the high voltage side where more monitoring is generally available. Demand measurements, which are not sent to the control room might be taken at ground mounted substations, while no monitoring is typically available on pole mounted substations or LV feeders, which are only temporarily monitored if there is a specific reason.

As power systems are transitioning to the next generation smart grids, new monitoring equipment is increasingly being installed at various locations of the distribution network. This equipment, which can be related to the ongoing deployment of smart meters, the increasing interest in LCTs, the active network management approaches that are adopted in larger areas of the network etc, leads to new streams of data in the distribution network. An example of the locations where monitoring equipment is expected to be installed is given in Figure 3.1, which shows the locations where enhanced network monitoring took place during Customer-Led Network Revolution (CLNR), which was the largest smart grid demonstration project undertaken by Northern Powergrid (NPG) [63].

Monitoring locations M1-M3 are at the HV level of the distribution network and refer to the monitoring of primary substations, HV feeders and HV connected customers as explained above. Monitoring locations M4-M8 refer the enhanced monitoring of secondary substations and various locations throughout the LV network that was trialled as part of the CLNR project. New data streams would result from the anticipated increase in monitoring systems. For example, based on Figure 3.1, additional measurements of current, voltage, power as well as power quality related data such as harmonics, flicker, unbalance etc would be expected to be generated at various locations of the distribution network (M4-M5). In addition, large amounts of data coming from the consumers would also be expected (M6-M8). When fully deployed, smart meters alone are expected to generate vast amounts of consumption data from the millions of electricity users throughout the country. The expected increase in data and information flow can be seen in Figure 3.2, which compares the traditional power system (left side of Figure

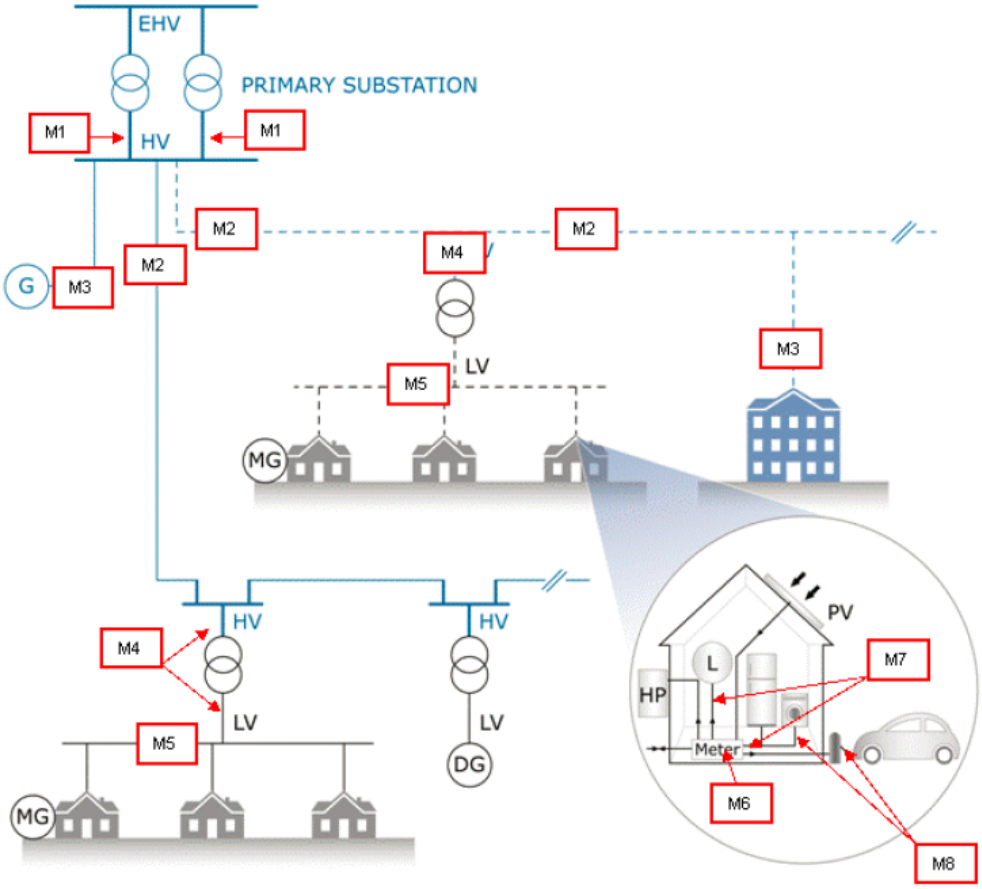


Figure 3.1: CLNR project monitoring locations [63].

3.2) and the anticipated flows following the roll-out of smart meters and the increase in distributed generation levels (right side of Figure 3.2) [64].

It is evident from Figure 3.2 that the replacement of traditional meters, which measure the electricity consumption and need to be manually read, with smart meters with additional functionalities as well as the increased uptake of LCTs and smart appliances will significantly change the data flows within the power system as well as the responsibilities of the DNOs. While, traditionally, the main responsibilities of DNOs have been to keep voltages and currents within limits ensuring uninterrupted supply, the use of smart meters will facilitate the participation of demand side in system balancing as they will have the capability to receive dynamic price signals and therefore allow

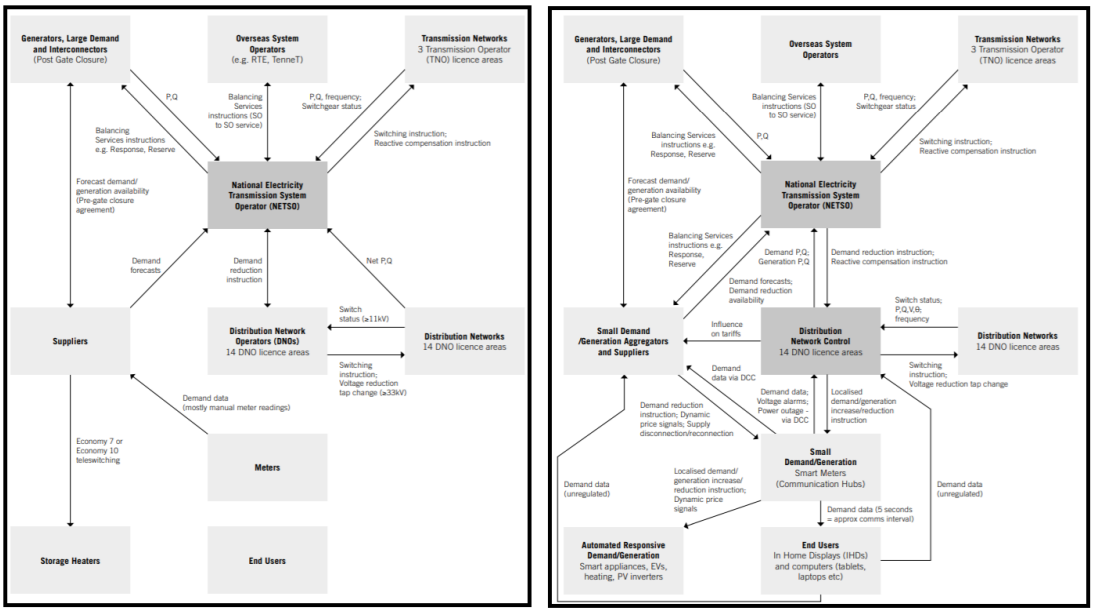


Figure 3.2: Data and information flow for the traditional (left side) and future power networks (right side). Adapted from [64].

consumers to change their consumption behaviour if required [64]. Smart meter data will make up a large part of the data generated in the smart grid, as a smart meter roll out to every home is expected in the next few years. According to [65] the volume of data generated by 1 million smart meters, which collect data every 15 minutes amounts to 2920 TB per year, which translates to 8 TB per day. Given the fact that there are 29 million homes in the UK [66], and assuming that all of them have smart meters collecting data every 15 minutes, this would lead to a data flow of 232 TB per day from smart meters alone.

The anticipated increase in data flows comes with a number of challenges for the DNOs, which are discussed in the next section.

3.2 Review and Assessment of Data Related Challenges

The data flows resulting from the increased monitoring of the distribution network have the potential to give an unprecedented view of the network and, with proper management, this data can help utilities significantly improve their services. To do so, however,

DNOs need to be aware of the challenges associated with the transformation of their networks and develop a data management policy that deals with these challenges.

The first and most obvious challenge is the volume of data that DNOs will need to be able to manage. The increased volumes of diverse data, which will have dynamic characteristics and will be collected at various resolutions, will require suitable collection and processing strategies to be developed by the DNOs [67]. As future networks will be heavily relied on data, there is no doubt that DNOs will eventually become a data-driven utility. To be able to accommodate the streams of data, and before considering the ways in which this data will be used to support network operation, a 'strong' grid needs to be established by upgrading their existing or developing a new robust Information and Communications Technology (ICT) infrastructure, which should be able to deal with the changing technologies and demand and should also guarantee secure access to data for every internal stakeholder that can extract value from this data [68], [69]. Another data related challenge is ensuring that data corresponding to confidential personal or organisational data are handled in a way that protects the user privacy and prevents unauthorised use of this data. Care should be taken, for example, when the consumer reaction to a dynamic price signal is investigated as, although prices are public, the consumption is confidential and one of the main concerns related to the use of smart meters is the profiling of usage patterns [70]. Apart from the new challenges associated with the increased data flows, there are a number of factors in the way existing data are currently handled by the DNOs. These factors can be broadly categorised in two groups [71]:

Issues related to the condition of data. Much of the data currently used for analysis is often of poor quality, which manifests itself mainly through gaps in the collected data. Data might be missing for various reasons. For example, this can be due to a failure in the monitoring system, data might have been collected but not stored, it might have not been collected at all or it can exist in a non-digital format.

Issues related to extracting value from data. Currently, much of the data is analysed from universities or third parties involved in joint projects, as the DNOs might not have

the expertise required for specific tasks. The highly restricted access to data for third parties limit their ability to extract value from data and, given the fact that DNOs currently are not the ones who analyse all the data collected in their networks, it leads to the value of data not being maximised, which in turn does not allow the network to operate as efficiently as it could.

Throughout the duration of this project, the interaction with DNOs as well as the condition of the DNO data that was used for the analysis presented in the next chapters of this thesis, revealed a number of challenges. These stemmed mostly from the way DNOs manage their existing data and confirmed some of the points made above. The challenges identified through the interaction with DNOs are discussed below.

Lack of online access to data: Due to security concerns, it was not possible to establish a VPN connection to the DNO's network that would allow for online access to the data. These concerns do not refer to the VPN technology, per se, but the DNO's reluctance to grant access to data from computers that were not connected to their network.

Limited access to data: The lack of online access mentioned above leads to a strong reliance on DNO employees who have other responsibilities as well and therefore less time to extract the required data. *Data on devices*: Although there is monitoring data available and the DNOs could benefit from its analysis, some of it remains on the monitoring devices as it requires someone to go on site to extract it. This takes a long time and is another factor that limits the amount of available for analysis data.

Existing monitoring equipment not being properly configured, meaning that data that could otherwise be used to improve network operation is not available.

Data synchronisation: DNOs were not always able to verify if different timestamps in seemingly correlated data were due to a synchronisation error.

Disparate datasets for the same application leading to increased requirements for data preprocessing.

Data stored in different platforms, making it difficult to identify and correlate all the necessary data. This can lead to data that is unsuitable for the application considered

to be provided by the DNO which can delay the analysis process.

DNOs do not always own the data: Third parties responsible for collecting and managing the data might refuse to grant access to it (or doing so might incur additional costs for the DNO).

While some of the challenges encountered throughout this project may be application specific or related to the engagement of certain people, the majority of these issues stem from the lack of a consistent data management for the DNO. Some recommendations on what would be necessary to consider when developing such a plan are given on the next section.

3.3 Recommendations

The 2019 Energy Data Taskforce's report [71], identified some of the above challenges and provided a set of recommendations that would allow network operators to fully utilise the data coming from the future decarbonised and decentralised grid, including:

1. The digitalisation of the energy system. This can be achieved by digitally enabling all infrastructure and assets and continuously improving the data quality as part of a long term data strategy.

2. Maximising the value of data by adopting an 'Open Data' approach while taking into account key concerns such as security and privacy and the treatment of confidential data.

3. Increased visibility of the data through the development of a 'Data Catalogue' that will include all data related to the energy sector, increase transparency across the sector and facilitate data sharing across organisations.

4. Coordination of asset registration that will simplify the energy asset registration process, which is currently very complex leading to reduced compliance with registration requirements.

5. The development of a 'Digital System Map' that would increase visibility of all energy infrastructure and assets and eventually lead to greater system resilience.

While all these recommendations will definitely help in maximising the value of data anticipated to flow in the future network, looking at the current state of data management at the DNO level, they seem unrealistic. In order to work towards this direction, a number of actions dealing with the factors that limit the current utilisation of existing data could be taken by the DNOs. The following actions could be considered.

Define their preferred strategy of data collection and processing based on their priorities. For example, collecting data at a fixed rate has the benefit of increased observability but also leads to increased storage requirements as more data will be generated. Alternatively, they can choose to collect data when an event is detected at the expense of observability [67].

Upgrade their ICT infrastructure in order to be able to cope with the anticipated data flows [69].

Use a common platform to store all different types of data collected in their networks, based on the principles outlined in IEC 61970 [72], which describes the Common Information Model (CIM) that was developed to allow interoperability between Energy Management System (EMS) applications and IEC 61968, which is the extension of the CIM in order to be used in distribution network management [73]. This will reduce the need to manage and maintain multiple systems and will give a clearer picture of what is monitored in the network. In addition, this will facilitate the data exchange and sharing between different departments of the same utility leading to better data utilisation as well as across organisations in the energy industry, which is one of the recommendations given by the Energy Data Taskforce in [71].

Unification of datasets. Storing similar data in the same format will reduce the data preprocessing requirements and will allow the application of models developed based on a specific dataset to other datasets when the same problem is investigated.

Remote transfer of all recorded data. This will reduce the need of people going to the

site to collect the monitoring data and increase the amount of data that can be actually utilised.

In the event that the data is analysed at a university or another company and not by the DNOs themselves, access to the required data for researchers and data scientists should be facilitated (e.g. in the form of online access to the data).

The above points regarding the improvement of data management procedures at the DNO level will need to be considered before deciding what the data analysis strategy will be. Although the current operational requirements can generally be met even without a consistent data management strategy, this will not be the case in the near future when the lack of such a strategy could make it impossible for a DNO to deal with the anticipated data flows and the increased responsibilities. The data analysis methodology proposed in the next chapter is a general framework that takes into account the challenges discussed above. Considering the current data related limitations, the methodology was demonstrated with case studies utilising the limited data available so it can offer immediate value to the DNO. If a data management and analysis strategy that can fully utilise the existing data is established now, DNOs could then extend this strategy to accommodate the future distribution network requirements.

Chapter 4

Data Analysis Methodology

The lack of a consistent data management strategy among DNOs was pointed out in the previous chapters. As the electrical grid is rapidly transforming to the new generation smart grid, new opportunities and challenges emerge. These are largely related to the amount of monitoring equipment and the vast amounts of data that are expected to be available for a distribution network. In order to be able to fully utilise the potential of the future smart grids, as well as to manage them effectively, the DNOs need to start developing a data management strategy that will deal with the tasks of data collection and storage, data processing and data analysis to support the operation of the distribution networks.

A detailed presentation of the methodology proposed in this thesis, is given in this chapter. This data analysis methodology involves the application of statistical analysis and machine learning algorithms on data coming from various sources. The first step of data analysis, before moving on to apply certain models, is the data cleansing and processing. The data analysis part of the proposed methodology consists of three stages: Characterising Network Behaviour, Anomaly Detection and Fault/Disturbance Prediction. The purpose of data pre-processing and that of each of the three analysis stages is explained in the next four sections, where the specific data analysis methods used in this thesis are also discussed. The metrics used for the assessment of the data analysis results are presented and explained in the last section of this chapter. The

work presented in the following chapters, demonstrate how applying this data analysis methodology can provide DNOs with valuable information that can be used to predict the occurrence of power quality events or faults and, therefore, improve the reliability of their network. The data analysis methodology proposed in this chapter can be summarised in Figure 4.1 below.

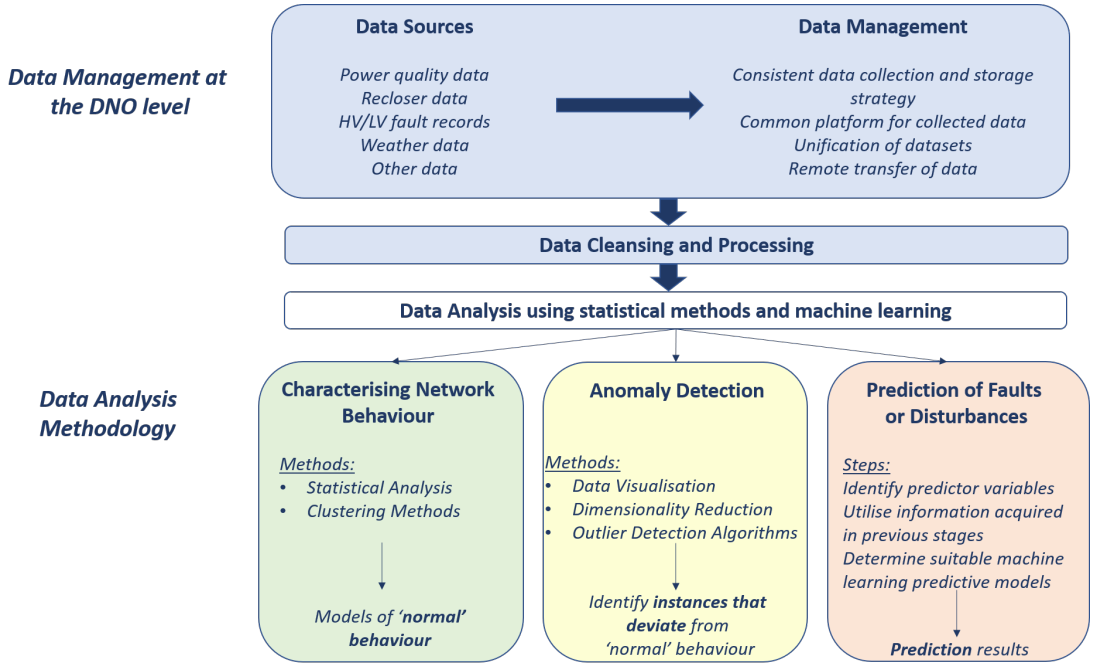


Figure 4.1: Decision support methodology.

The data sources shown in the above diagram correspond to the data used throughout this thesis. The "Other data" mentioned in the data management block indicates that a similar approach can be applied to other data sources as well.

The methodology illustrated in Figure 4.1 is demonstrated with a series of short and more detailed case studies presented in the next chapters. More specifically, the three short case studies of Chapter 5 provide examples of data analysis falling into the first two stages of the methodology, as statistical analysis is used to identify typical behaviour ('Characterising Network Behaviour') and then cases that do not follow the expected behaviour are detected using data visualisation ('Anomaly Detection'). Next, a more detailed case study focusing on the 'Prediction of Faults or Disturbances' stage of the

methodology is presented in Chapter 6, where the impact of weather on the occurrence of fault is investigated. Finally, Chapter 7, which studies the impact of load behaviour on the occurrence of power quality disturbances, provides an application that involves all three stages of the data analysis methodology. In the first stage, the typical load profiles per substation are identified ('Characterising Network Behaviour'), while in the second stage dimensionality reduction and visualisation is used to identify a possible relation between load behaviour and event occurrence ('Anomaly Detection). This identified relation is then confirmed in the event prediction stage ('Prediction of Faults or Disturbances'). The machine learning and statistical methods used for the above case studies are discussed in the next sections of this chapter.

4.1 Data Preparation and Pre-processing

Data preparation and preprocessing is one of the most important tasks involved in data analysis. A number of different steps need to be taken in order to transform the raw data into a form that is suitable for further analysis, e.g the application of machine learning models. This process involves a number of steps that range from the initial exploration of data to the point where the data can be fed to the models. [74] These steps include dealing with inconsistent, duplicate or missing values (e.g. imputation) as well as transforming the data to a suitable format before applying the machine learning models (e.g standardisation/normalisation), among other things. The main steps of data preparation and preprocessing are discussed in this section, where the way that these steps have been used in this thesis is also pointed out.

Data exploration: One of the first steps of data analysis is the initial exploration of data to understand the data and its characteristics and reveal as much information as possible. This step can help define the problem that needs to be addressed and provides insight on how it should be addressed. Data exploration is usually part of the first stage of this methodology and, therefore, a more detailed description of data exploration is given in the next section, where the "Network Behaviour Characterisation" is discussed.
Dealing with missing values: Missing data can be identified through the initial data

exploration and the way the missing data is handled can be crucial for the next steps of the analysis. Depending on the missing data and the application, the data examples with missing values can be discarded or they can be used for the analysis (after the empty values have been replaced by other values in a process called imputation). There are various ways to perform imputation of missing data. For example, the missing value can be replaced with the mean value of the relevant variable, with the most common value, with the value of the previous data point etc.

Dealing with categorical values: Machine learning models generally require numerical values as inputs. However, there are categorical variables, which need to be dealt with in order to proceed with the analysis. For example, a categorical variable which can take n values, can be represented by n new variables taking the values 0 and 1 to indicate the absence or the presence of the event.

Feature scaling: Another common requirement of most machine learning methods is that the input data needs to be scaled. It is typical for some methods to assume Gaussian distribution with zero mean and unit variance, therefore they might not perform as they should if this requirement is not met. Throughout this thesis, the data has been scaled to fulfil this requirement.

Training and testing datasets: Once some or all of the previous steps have been done and the resulting dataset has no missing or inconsistent values, and before applying a machine learning model, this dataset needs to be split into training and test set. In this thesis, 80% of the data has been used to train the predictive models, while the remaining 20% has been used to test their performance.

In addition to the above steps, *Dimensionality reduction* and *Feature extraction* can also be used as preprocessing steps to support subsequent activities. Dimensionality reduction refers to the process of reducing the number of features and can be achieved by feature selection, which keeps a subset of the original features or feature extraction, which creates new ones. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two commonly used methods for feature extraction, however, this was not how they were used in this thesis. PCA was used in Chapter 7 to visualise the clusters obtained from the clustering model, while LDA was used in Chapter 6 as

a predictive model.

Some of the data preprocessing steps discussed in this section have been incorporated in the three data analysis stages detailed in the next chapters.

4.2 Characterising Network Behaviour

The purpose of this section is to investigate the available data in order to identify the normal operating conditions of the network with respect to the problem that is being examined and to develop representative models that can be used for further analysis. The characterisation of network behaviour is the first stage of the proposed methodology and represents the conditions under which the network is used for certain applications. This process helps a DNO understand each application's requirements from the network, which in turn will help them identify potential network maloperation or unexpected events when the collected data indicate different operating conditions from those that have been identified as "normal". This stage involves the initial exploration of available data through statistical analysis or the generalisation of behaviour using clustering methods. The specific methods that have been used in the case studies discussed in this thesis are presented below.

4.2.1 Statistical Analysis and Exploration of Data

In his 1977 book called "Exploratory Data Analysis", John Tukey described what an exploratory data analysis approach is and how it could be used when dealing with data [75]. Rather than giving a fixed set of techniques, this book sets out the principles of how data analysis should be carried out. In other words, it explains how an analyst should handle a given dataset, what they should look for and how to look for it, as well as how to interpret the findings. The purpose of such an analysis is to reveal as much information contained in a dataset as possible, before going into a deeper and more specific type of analysis. This includes identifying hidden patterns within the data, finding relations between variables and extracting potentially important variables based on the observed behaviour, as well as detecting outliers and unusual behaviour

within the data. As the datasets are usually in the form of tables containing values for several numerical and categorical variables, this initial exploration process, is not always straightforward just by looking into the data. Therefore, the techniques used for exploratory data analysis are mainly graphical representations of either the raw data or simple statistics such as the mean, median, standard deviation etc and include simple plots of variables, histograms, scatter plots, box and whisker plots, probability distributions etc. The visual perception and the inherent pattern recognition capabilities that humans possess are very important in exploratory data analysis. As discussed in the previous section, the initial exploration of data is part of data preprocessing.

From what was discussed above, it is evident that exploratory data analysis using statistical methods is, generally, one of the first steps of analysing data in order to get an insight of what this data can tell. With respect to the methodology presented in this thesis, the process described above covers parts of both the "Characterising Network Behaviour" and "Anomaly Detection" stages. Regarding the first stage, scatter plots, histograms and box and whisker plots have been widely used during the initial exploration of data for each of the case studies discussed in this thesis, although not all of them are presented in the next chapters where the case studies are discussed. For the stage of "Anomaly Detection", the dimensionality reduction and visualisation techniques that were used throughout this thesis to reveal anomalies in the data are discussed in Section 4.2.

4.2.2 Clustering of Data Using Gaussian Mixture Models

The stage of network behaviour characterisation for the case study presented in Chapter 7, involved the identification of recurrent load profiles which could then be used to investigate the impact of load behaviour on the occurrence of power quality events. The method used as a clustering model to group the load data and identify these representative load profiles was a Gaussian Mixture Model (GMM), which is explained in this section.

In GMM, it is assumed that all data points to be clustered come from a linear combi-

nation of a finite number of Gaussian distributions. The distribution parameters are learned using the Expectation-Maximisation (EM) algorithm which enables a probabilistic clustering to be performed by the trained model. Each data point has a probability of belonging to each of the obtained distributions and the final assignment is performed by selecting the component (cluster) with the highest probability [76]. This process is explained in more detail below.

A Gaussian Mixture is a combination of k Gaussian distributions $g_1, g_2, ..., g_k$, where k is the number of clusters. Each of these distributions is characterised by a mean (μ), a variance (σ^2) and a weight (or mixing proportion (π)), which define the centre, the width and the contribution of the Gaussian to the mixture respectively. The probability density function of a Gaussian distribution can be determined by its mean and variance as can be seen in equation 4.1.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{4.1}$$

The mixing proportions $\pi_1, \pi_2, ...\pi_k$ (where k is the number of Gaussians in the mixture) are probabilities and their sum amounts to 1, as shown in equation 4.2.

$$\sum_{j=1}^{k} \pi_j = 1 \tag{4.2}$$

When using GMMs to perform clustering on a group of data points, each point is assumed to have come from a Gaussian but which data points came from which Gaussian is not known in advance. In order to cluster the data, the above parameters, and therefore the k Gaussians need to be determined. This is done using the EM algorithm, which starts with a random initialisation of the k Gaussians. Given the group of k Gaussians, the probability of each data point to belong to each of these Gaussians is checked and then the Gaussian that each point is more likely to belong to is identified and the data points are assigned to the relevant clusters (Expectation step). The probability that the point x_i came from the Gaussian g_1 is given by the equation

4.3

$$P(x_i|g_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i-\mu_1}{\sigma_1})^2} \tag{4.3}$$

This is calculated for all Gaussians and then the Bayesian posterior probability $g_{1,i}$ (which is the conditional probability after the relevant evidence has been taken into account) is calculated for each point x_i using the above probability for each Gaussian as shown in equation 4.4.

$$g_{1,i} = P(g_1|x_i) = \frac{P(x_i|g_1)P(g_1)}{P(x_i|g_1)P(g_1) + P(x_i|g_2)P(g_2) + ... + P(x_i|g_k)P(g_k)} \tag{4.4}$$

The probabilities $P(g_1), P(g_2), ..., P(g_k)$ are the prior probabilities and reflect what proportion of the data points are described by each of the Gaussians. These are usually considered to be equal in the first step of the EM algorithm.

The posterior probabilities are compared and the data point x_i is assigned to the Gaussian with the highest posterior probability. After this assignment, the data points covered by each Gaussian are used to re-estimate the means (μ_j) and variances (σ_j) of the k Gaussians, taking into account the data points that were assigned to each Gaussian and their respective probabilities, as shown in equations 4.5 and 4.6 respectively (Maximisation step).

$$\mu_1 = \frac{g_{1,1}x_1 + g_{1,2}x_2 + ... + g_{1,n}x_n}{g_{1,1} + g_{1,2} + ... + g_{1,n}} \tag{4.5}$$

$$\sigma_1^2 = \frac{g_{1,1}(x_1 - \mu_1)^2 + g_{1,2}(x_2 - \mu_1)^2 + ... + g_{1,n}(x_n - \mu_1)^2}{g_{1,1} + g_{1,2} + ... + g_{1,n}} \tag{4.6}$$

Equations 4.5 and 4.6 calculate the mean (μ_1) and variance (σ_1^2) of Gaussian g_1 from the n data points that have been assigned to this Gaussian, using the posterior probabilities calculated in 4.4.

The process involves many iterations and it is repeated until convergence. Then, the optimal values for the mean, variance and weight for each of the Gaussians are those of the last iteration.

The above procedure refers to the case of 1-dimensional data, where only 1 variable is considered but is similar when the analysis involves high-dimensional data. In the work presented in this thesis, GMMs are used to cluster high-dimensional data. More specifically, each daily load profile is described by 24 variables, which are the average current measured at each hour of the day. In the case of clustering of high-dimensional data, each data point with d attributes is assumed to have come from a group of k sources $c_1, c_2, ..., c_k$, which are Gaussian distributions in a d-dimensional space. The process followed in this case is similar but instead of the mean and variance, a mean vector and the covariance matrix, which shows the covariance between the d attributes, need to be calculated. The application of EM algorithm to fit a multivariate GMM follows the same steps with those described above. The probability that each instance $\vec{x_i}$ belongs to each of the sources is calculated as shown in equation 4.7, where the probability for source c_1 is calculated.

$$P(\vec{x_i}|c_1) = \frac{1}{\sqrt{2\pi|\Sigma_{c_1}|}} \exp\{-\frac{1}{2}(\vec{x_i} - \vec{\mu_{c_1}})^T \Sigma_{c_1}^{-1} (\vec{x_i} - \vec{\mu_{c_1}})\} \tag{4.7}$$

where $\vec{x_i} = [x_1, x_2, ..., x_d]$ is the ith d-dimensional data point, $\vec{\mu_{c_1}}$ is the mean vector and Σ_{c_1} the covariance matrix for source c_1.

Similarly to the univariate case described above, the posterior probability for each of the multi-dimensional data points $\vec{x_i}$ to belong to source c_1 is calculated as follows:

$$P(\vec{x_i}|c_1) = \frac{P(\vec{x_i}|c_1)P(c_1)}{\sum_{j=1}^{k} P(\vec{x_i}|c_j)P(c_j)} \tag{4.8}$$

Consequently, the mean and covariance for source c_1 are calculated using equations 4.9 and 4.10 respectively, where $\mu_{c_1,j}$ is the mean of Gaussian c_1 for attribute j across all data points that belong to this source and $x_{i,j}$ indicates the jth attribute of the ith

data point.

$$\mu_{c_1,j} = \sum_{i=1}^{n} (\frac{P(c_1|\vec{x_i})}{nP(c_1)}) x_{i,j} \tag{4.9}$$

$$(\Sigma_{c_1})_{j,k} = \sum_{i=1}^{n} (\frac{P(c_1|\vec{x_i})}{nP(c_1)}) (x_{i,j} - \mu_{c_1,j})(x_{i,k} - \mu_{c_1,k}) \tag{4.10}$$

In contrast to methods such as k-means clustering, which perform a hard clustering on data, meaning that each point is assigned to one cluster only with no uncertainty measure associated with it, GMMs perform a soft clustering, where each data point belongs to all clusters with a different probability and is assigned to the one with the highest probability. It is worth noting that k-means is essentially a simplified version of GMM, as it can be obtained as a particular non probabilistic limit of EM, when it is applied to Gaussian mixtures [77]. Two additional parameters that need to be determined when using a GMM are the 'covariance type' and the 'number of components'. Covariance type controls the shape and position of each cluster and is reflected by the covariance matrices of the GMM's components. The number of components (or clusters) k can be known in advance or it may need to be inferred from the data [76].

4.3 Anomaly Detection

Anomaly detection is the identification of unusual or rare events that are reflected by data points which seem to behave in a very different way than the majority of the analysed data. Anomaly detection is similar to novelty detection but it differs in that the anomalies are unusual behaviour of data that needs to be avoided, while novelties reflect previously unobserved behaviour that can be generally considered as an aspect of normal behaviour after they have been detected [78]. Both types of outliers described above (anomalies and novelties) are covered in the 'Anomaly Detection' stage of this methodology, which is the subject of this section.

In the context of power systems, these anomalies can be related to unusual operation which does not necessarily pose a threat to the network or they could be related to a truly problematic network operation. For example, they can reflect the consumers and/or the network's response to a rare event, i.e. unusual consumption patterns for a specific time of day or season due to a rare or unexpected event. However, in the second and more serious case, the observed anomalies might reflect actual faults or conditions that have an impact to the quality and security of supply. In both cases, however, anomaly detection can help DNOs gain an insight on their network and being able to understand the conditions associated with these anomalies can be beneficial. When DNOs are aware of typical harmless anomalies that may occur under certain conditions, they can take the necessary measures to minimise disruption on the network, while in the case of anomalies that reflect genuine faults, they may be able to identify certain patterns that would help them to detect evolving faults or conditions that might lead to a fault.

The 'Anomaly Detection' stage of the proposed methodology might not always follow the 'Network Behaviour Characterisation' stage as, in many cases, anomalies can be detected during the exploratory data analysis that is part of the first stage of the methodology and was described in the previous section. Throughout this thesis, techniques involving data visualisation were used to identify unusual behaviour within the data. Visualisation of data involves both the use of simple plots of the raw data, by taking single variable plots or those that can reveal the possible relationship between variables, as well as visualisation of high-dimensional data into a low-dimensional space after that data has been processed, using dimensionality reduction techniques. The dimensionality reduction and visualisation techniques that were considered in the case studies presented throughout this thesis are Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE), which are described below.

4.3.1 Principal Component Analysis

Principal Component Analysis (PCA) was developed by K. Pearson in 1901 [79] and is one of the most commonly used dimensionality reduction methods. PCA works by transforming the data into a set of orthogonal principal components, which are linear combinations of the initial variables and contain as much of the information contained in the data as possible. The first principal component is selected in such a way that it accounts for the highest possible variability in the data and each succeeding orthogonal component contains as much of the remaining variance as possible. This transformation allows the representation of the initial high-dimensional data on a space of lower dimensions formed by uncorrelated variables which preserved as much of the information contained in the initial correlated variables.

The reason why the dimension with the highest variance is selected for the principle components is that these dimensions preserve the relative distances between the original data points. The procedure to select the principal components for a d dimensional data set $X = [x_1, x_2, ..., x_d]$ is described below.

First, the data is standardised by subtracting the mean μ from all attributes and then the covariance matrix Σ is calculated using the equations 4.11 and 4.12.

$$cov(a, b) = \frac{1}{n} \sum_{i=1}^{n} x_{ia} x_{ib} \tag{4.11}$$

$$var(a) = \frac{1}{n} \sum_{i=1}^{n} x_{ia}^2 \tag{4.12}$$

Equation 4.11 gives the covariance between two attributes a and b and is used to fill the non-diagonal elements of the covariance matrix Σ, while the diagonal elements show the variance of each of the attributes and their values are calculated using 4.12. As the data have been centred, they have zero mean and this is why the mean is not subtracted in the above equations. Now that the covariance matrix Σ is calculated, the m principal components of the high dimensional data are the m eigenvectors of Σ with

the largest eigenvalues. The relation between the eigenvectors $\mathbf{e}$ and eigenvalues λ of a covariance matrix Σ is given in 4.13.

$$\Sigma\mathbf{e} = \lambda\mathbf{e} \tag{4.13}$$

The eigenvalues λ_i are calculated by solving

$$\det(\Sigma - \lambda I) = 0 \tag{4.14}$$

and then are used to calculate the eigenvectors from equation 4.13. The 1^{st} principal component (eigenvector) is the one that corresponds to the biggest eigenvalue, the 2^{nd} is that with the next bigger eigenvalue etc. The initial data X are then projected to the principal components, which are the axes of the low dimensional space. For a data visualisation on a m dimensional space (where $m << d$), the m principal components are selected and the resulting data $X\prime = [x_1\prime, x_2\prime, ..., x_m\prime]$ are calculated using:

$$\begin{bmatrix} x_1\prime \\ x_2\prime \\ \vdots \\ x_m\prime \end{bmatrix} = \begin{bmatrix} (\vec{x} - \vec{\mu})^T \vec{e_1} \\ (\vec{x} - \vec{\mu})^T \vec{e_2} \\ \vdots \\ (\vec{x} - \vec{\mu})^T \vec{e_m} \end{bmatrix} = \begin{bmatrix} (x_1 - \mu_1)e_{1,1} + (x_2 - \mu_2)e_{1,2} + \cdots + (x_d - \mu_d)e_{1,d} \\ (x_1 - \mu_1)e_{2,1} + (x_2 - \mu_2)e_{2,2} + \cdots + (x_d - \mu_d)e_{2,d} \\ \vdots \\ (x_1 - \mu_1)e_{m,1} + (x_2 - \mu_2)e_{m,2} + \cdots + (x_d - \mu_d)e_{m,d} \end{bmatrix} \tag{4.15}$$

4.3.2 t-distributed Stochastic Neighbour Embedding

One of the visualisation methods used widely in this thesis is t-Distributed Stochastic Neighbour Embedding (t-SNE), which is a very effective method for data intensive tasks as demonstrated in [80] via a comparison with other related techniques. Given its good performance on high-dimensional data sets for applications in various sectors, it was decided to investigate how t-SNE would work with high-dimensional power system data and whether it could be used to identify instances suggesting unusual behaviour

of the network. This dimensionality reduction method was developed in 2008 by L. van der Maaten and G. Hinton and is described in [80]. It provides a way of visualisation of high dimensional data $X = \{x_1, x_2, ..., x_n\}$ in a two or three dimensional map $Y = \{y_1, y_2, ..., y_n\}$, while preserving the local structure of the data. This allows the high dimensional data to be easily presented to the user using a scatter plot. t-SNE is an updated version of an earlier dimensionality reduction technique called SNE but it differs in two points. The process of how SNE performs mapping of high dimensional data points to the low dimensional space is briefly described below, followed by the two updates employed by t-SNE.

Each point x_i in the high dimensional space has a Gaussian distribution centred at x_i. First, SNE converts the Euclidean distances between two points in the high dimensional space into conditional probabilities that represent similarities. Therefore, the similarity between two points x_i, x_j is the conditional probability that x_j is the closest neighbour of x_i, provided that the neighbours are chosen proportionally to their probability density under a Gaussian centred at x_i. The expression used to calculate this conditional probability (similarity) is:

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{4.16}$$

This means that the conditional probability (or similarity) between two nearby points is high, while for distant points it will be very small. For the low dimensional representation of this data, SNE starts by placing the data points at random locations on the low dimensional map. The similarity between points y_i and y_j is calculated in a similar way but, in this case, the variance of the relevant Gaussian distribution is set to $\frac{1}{\sqrt{2}}$

$$q_{j|i} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq i} \exp(-||y_i - y_k||^2)} \tag{4.17}$$

In order to have an accurate representation of the high dimensional data to the low di-

mensional space, these two probabilities need to be equal and to do this SNE minimises the mismatch between the conditional probability distributions P_i and Q_i, which are taken over all other data points given x_i and y_i respectively. Mathematically, the sum of the Kullback - Leibler (KL) divergence (which is a measure of the difference between two probability distributions) over all data points is minimised as shown in equation 4.18.

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{4.18}$$

The above method (which is the SNE method mentioned earlier in this section) has two drawbacks: the first is related to the cost function C which tends to retain the local structure of the low dimensional representation of the data (which is randomly placed in the beginning of the process); the second is that it tends to represent high dimensional data points of moderate distance, with low dimensional points which are very close to each other (crowding problem). To deal with these problems, t-SNE uses a symmetric cost function and a Student t-distribution to convert the pairwise distances to probabilities in the low dimensional space (the Gaussian distribution is still used for the high dimensional space).

Symmetric cost function

Instead of minimising the sum of KL divergences for the conditional probability distributions P_i and Q_i, t-SNE minimises a single KL divergence between the joint probability distributions P and Q in the high and low dimensional spaces respectively, as shown in equation 4.19.

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4.19}$$

In the above equation, $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji} \; \forall \, i, j$, hence it is a symmetric cost function that is minimised by t-SNE. The pairwise similarities for the high dimensional space p_{ij} used in equation 4.19 are calculated using:

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2/2\sigma^2)} \tag{4.20}$$

The relevant probabilities for the low dimensional space are calculated using the Student t-distribution discussed below.

Student t-distribution

A similar way is used to convert distances into probabilities in the low dimensional space, but instead of a Gaussian distribution, a student-t distribution is used. The student t-distribution is used instead of a Gaussian in the low dimensional space because it has a heavier tail than the Gaussian and this allows moderate distances in the high dimensional space to be modelled by larger distances (than the distances if Gaussian was used) in the low dimensional space. By doing this, the concentration of moderately dissimilar data points in the same area is avoided and thus much of the local structure of the high dimensional space is preserved in the low dimensional map. Using a t-distribution with 1 degree of freedom, the probabilities q_{ij} are:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}} \tag{4.21}$$

Finally, t-SNE tries to minimise the mismatch between the probability distributions as explained above. If the two probability distributions are equal, that means that the high dimensional points have been correctly mapped to the low dimensional space.

4.4 Prediction of Faults or Disturbances

The third stage of the proposed methodology involves identifying the predictor of faults or disturbances in the distribution network and is the main topic of both case studies presented in Chapters 6 and 7. The purpose of this stage is to utilise the information that has been acquired during the first and second stages in order to develop a prediction model using machine learning. The selection of the more suitable machine learning techniques and the input variables for each application are part of this stage.

4.4.1 Data Pre-processing

The preprocessing steps relevant to this stage involve the selection of features that were used as inputs to the predictive models, their scaling and the selection of training and test sets. For the case of weather-related fault prediction discussed in Chapter 6, the selection of features involved the initial exploration of the available weather data, the cleaning of the relevant datasets and the calculation of sums and differences that were used as inputs alongside the measured values. For the case of power quality event prediction discussed in Chapter 7, the selection of features involved looking into the clusters obtained from the previously applied GMM and their characteristics (such as the weight of each cluster in the GMM) as well as the visualisation of the clustering results to reveal if there is a relation between the changing states of load and event occurrence. The scaling of features and the splitting of the relevant datasets into training and test sets were the same for both prediction cases. The features were standardised so that they have a zero mean and unit variance, while 80% of the data was used for the training of the models and 20% for testing. The training and test datasets were selected in a way that the ratio of the classes in the original dataset was maintained.

4.4.2 Predictive Models

A number of different classification methods with diverse underlying decision surfaces to cover a wide range of applications are presented in this section. All the classification techniques discussed below were compared in order to identify the best performing method for the prediction of weather-related faults in Chapter 6. Out of these techniques Gradient Boost and Linear Discriminant Analysis were found to perform better, for the cases of LV and HV faults respectively. For the case study presented in Chapter 7, only Gradient Boost was considered. The working principles of the above mentioned classifiers are presented in this section, which also includes a brief description of the rest of the techniques considered. It is worth noting that all of the machine learning methods that are presented in this section have been used as classification methods for the prediction of weather related faults, even though some of them can be used for

other purposes such as dimensionality reduction, feature extraction etc.

Gradient Boost

In machine learning, boosting is a method that combines many simple models in an ensemble that performs better than the individual models, by sequentially applying a weak classifier to repeatedly modified versions of the data. The first successful implementation of boosting was the AdaBoost algorithm, which is short for adaptive boosting. Gradient Boost (GB), which is a generalisation of AdaBoost, uses multiple single predictors (often trees or rules) to make very simple, broad classifications and then weighs outputs from these according to their expected error into an overall classifier that has greater predictive power than any one of the constituent predictors [81]. The classification process starts with a very simple model (e.g. a decision tree), which is a weak classifier, meaning that it produces predictions which are only slightly better than guessing. Then subsequent models are used to predict the error made by the model so far. The models, which are trained sequentially, focus on the difficult to predict data examples. The objective of these classifiers is to minimise the loss, which is the difference between the actual and the predicted class value of a training example. To minimise this loss, this method uses gradient descent. Each model is evaluated and given a weight and the overall prediction of the classifier is given by a weighted sum of the collection of models. A brief description of AdaBoost is given below, followed by the improvement introduced by Gradient Boost.

AdaBoost starts by setting the observation weights $w_i = 1/N$, where $i = 1, 2, ..., N$ and N is the number of the training samples (x_i, y_i). At each step m (where $m = 1, 2, ..., M$), AdaBoost fits a classifier $G_m(x)$ to the training with weights w_i. Then, the error err_m is calculated and it is used to compute the parameters a_m which show the contribution of each classifier G_m. When these parameters are known, they are used to adjust the observation weights w_i. The classifier weights a_m are higher for the more accurate classifiers, while the observation weights w_i are higher for the misclassified examples. Therefore, the subsequent classifiers focus on the difficult to predict examples. The final prediction is obtained by a weighted majority vote on the results of the M classifiers

as shown in equation 4.22:

$$G(x) = sign(\sum_{m=1}^{M} a_m G_m(x)) \tag{4.22}$$

Although this method was a great improvement compared to a single decision tree, it it is limited to simple classification tasks and works for a specific type of loss function (exponential loss). As mentioned earlier, Gradient Boost is a generalisation of AdaBoost and works with any arbitrary differentiable loss function, making it more suitable for a wider range of applications.

Unlike AdaBoost, which assigns high observation weights to the difficult to predict examples, Gradient Boost identifies these examples using residuals, which are calculated in each iteration m (and for each class k) using equation 4.23. The difficult to predict data examples are identified by large residuals.

$$r_{k,im} = -\left[\frac{\partial L(y_i, f_k(x_i))}{\partial f_k(x_i)}\right]_{f_k = f_{k,m-1}} \tag{4.23}$$

where $L(y, f(x))$ is the loss function.

The residuals are then used to train a weak classifier $h_{km}(x)$ which is multiplied by a multiplier γ_{km}. This is calculated using:

$$\gamma_{km} = \arg\min_{\gamma}(\sum_{i=1}^{N} L(y_i, f_{k,m-1}(x_i) + \gamma h_{km}(x_i)) \tag{4.24}$$

In the above equation, gradient descent is used to find the γ that minimises this expression. The model is then updated to

$$f_{k,m}(x) = f_{k,m-1}(x) + \gamma_{km} h_{km} \tag{4.25}$$

This process is repeated K times at each iteration m, one for each class and the final model is given by K different (coupled) tree expansions f_{kM}, where $k = 1, 2, ..., K$,

which produce the probabilities that a data point x_i belongs to each of the K classes, as explained in [81].

Linear Discriminant Analysis

To make predictions, Linear Discriminant Analysis (LDA) as the name suggests, uses a linear decision boundary implied by the intersection of probability distributions representing difference classes [81]. A linear separation of the data is achieved when the data points are separated by class using a line or a hyperplane in the d dimensional input variable space. Similarly to PCA, LDA also finds a linear combination of input variables. The high dimensional data points are then projected on this eigenvector (as in PCA) and a hyperplane perpendicular to this vector is used as the linear decision boundary used for data classification (LDA can also be used for dimensionality reduction like PCA). In LDA, two simplifying assumptions about the data are made: (i) the data is assumed to follow a Gaussian distribution and (ii) the inputs of every class have the same covariance. A brief explanation of LDA for the case of binary classification of a high dimensional data point $\vec{x} = [x_1, x_2, ..., x_d]$, where each data point $\vec{x}$ belongs to one of two classes, namely $y = 0$ or $y = 1$, is given below.

To perform classification, LDA assumes that the conditional probability density functions of the two classes $p(\vec{x}|y = 0)$ and $p(\vec{x}|y = 1)$ follow a normal distribution. The mean vectors and covariance matrices of these distributions are assumed to be $(\vec{\mu_0}, \sum_0)$ and $(\vec{\mu_1}, \sum_1)$ respectively. Based on the above, a data point $\vec{x}$ belongs to the second class (here $y = 1$) if the log of the likelihood ratio is bigger than some threshold T, so that:

$$(\vec{x} - \vec{\mu_0})^T \Sigma_0^{-1} (\vec{x} - \vec{\mu_0}) + \ln |\Sigma_0| - (\vec{x} - \vec{\mu_1})^T \Sigma_1^{-1} (\vec{x} - \vec{\mu_1}) - \ln |\Sigma_1| > T \tag{4.26}$$

After the application of the second assumption of LDA, which is that the classes have common covariance matrices ($\Sigma_0 = \Sigma_1 = \Sigma$), the above expression is simplified and the decision criterion becomes a threshold on the dot product

$$\vec{w} \cdot \vec{x} > c \tag{4.27}$$

where $\vec{w} = \Sigma^{-1}(\vec{\mu_1} - \vec{\mu_0})$ and c is some threshold constant given by

$$c = \frac{1}{2}(T - \vec{\mu_0}^T \Sigma^{-1} \vec{\mu_0} + \vec{\mu_1}^T \Sigma^{-1} \vec{\mu_1}) \tag{4.28}$$

In terms of classification of a data point $\vec{x}$ to one of the two classes, it is determined by which side of a hyperplane (that is perpendicular to $\vec{w}$) this point is located on. The threshold c determines the location of the above hyperplane.

Additional Classification Methods

The classification performed by a number of different classification methods on three sets of synthetic data is shown in Figure 4.2, which was adapted from [82]. The scatter plots below each classifier illustrate the different decision boundaries, while the value shown at the bottom of the plots is the resulting classification accuracy. The data points used for training of the classifiers are indicated by the use of solid colors, while the semi-transparent data points form the test data.

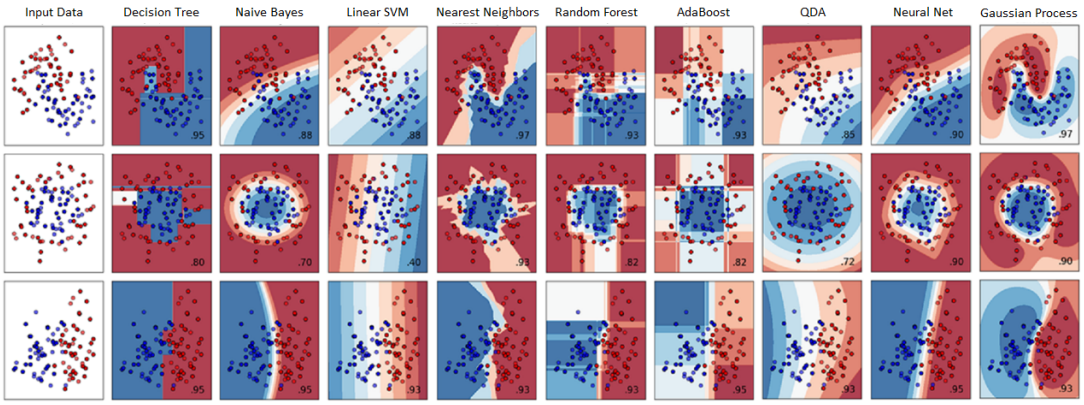


Figure 4.2: Classifier comparison [82].

Figure 4.2 does not include all the classifiers discussed in this section but it is a good illustration of the various types of decision boundaries used for data classification. A

brief description of the classifiers that were considered in Chapter 6, and were not discussed earlier in Section 4.3, is given below.

Classification And Regression Trees (CART) are one of the most established and common techniques used for data classification. A decision tree is constructed using historical data, where each data point belongs to one of two or more classes. That tree, which is represented by certain rules, is then used to predict the class of new, previously unseen, data examples. While functionally limited, CART is simple enough to produce a classifier that can be interpreted as a decision tree [83].

Naive Bayes (NB) works by isolating classes within data using joint probability structure, as the probabilities of each feature for each class are multiplied and then the class with the highest probability is selected. The 'naive' part of the name comes from the fact that observations are assumed to be completely independent thus making the decision boundaries for classification almost as simple as those of CART but for the fact that coinciding observation probabilities are taken into account [84].

Logistic Regression (LR) follows a similar approach but uses a logistic sigmoid at the decision boundary to accommodate arbitrary closeness of a data point within the range of a class [77].

In order to perform data classification, ***Support Vector Machines (SVM)*** construct a hyperplane, which divides the data points. The classification is considered better when the distance from the hyperplane of the nearest example from each class is larger. In the case of high dimensional data a set of hyperplanes are constructed. To deal with data points that are not linearly separable, SVMs map the input variables into a high dimensional space to provide greater discriminative power [85].

k-Nearest Neighbour (k-NN) is another very common classification technique. The idea behind this method is that an unclassified data point is assigned to the class that corresponds to the majority of its k nearest neighbours. In k-NN, the labels of new data points are assigned after the neighbouring points are determined using a metric such as the Euclidean distance [86].

In Bagged Trees (BT), bagging, which is another method for creating an ensemble from a collection of simpler models, is applied on a decision tree classifier. This means that the classifier is trained on random subsets of the original training dataset and the final prediction is constructed by aggregating the individual predictions. In bagging, the samples that form the random subsets are drawn with replacement [87]. ***Random Forest (RF)*** is a modified version of bagged trees, where many trees with reduced correlation are developed and then averaged [81].

Quadratic Discriminant Analysis (QDA) is similar to LDA which was explained above but it differs in that it relaxes the assumption that classes have the same covariance structure, resulting in class decision boundaries that are quadratic [81].

Multi-Layer Perceptrons (MLP) allow arbitrary nonlinear functions to be modelled either as regressors or classifiers. This is enabled by layers of hidden weights that are tuned to map model inputs to outputs through a process called Backpropagation [88].

Gaussian Process Classification (GPC) is a very flexible classification model that can take a number of decision surface shapes. It achieves this by modelling decision boundaries as a Multivariate Gaussian distribution with position dictated by its mean and decision threshold by is covariance function - this can be specified to take a different form along the length of the decision boundary [89].

The trade-off to be made when comparing different classification methods is one of complexity versus generalisation: the relation between fault occurrence and complex weather phenomena that is explored in Chapter 6 may not be captured by a simple classification boundary. However, closely fitting a classification boundary to very specific weather conditions is also undesirable as the classifier will capture too few eventualities - this phenomenon is referred to as overfitting [88].

The machine learning methods discussed in Sections 4.1-4.3 are those used in the case studies presented later in this thesis and have been chosen among various clustering, dimensionality reduction and classification techniques. This does not imply that these

are the only methods that can be used for relevant applications and the reasoning for the selection of these specific methods is the following. The existing literature regarding the clustering of load profiles shows that k-means clustering is the most common method and has been widely used for this kind of data, followed by GMM, which have also been proven to be successful in relevant applications. As k-means clustering is essentially a simplified version of GMM, it was decided to use the latter as it gives a better insight of the clustering by providing the probability of the labels assigned by the clustering process. In addition, a comparison among various distributions has been performed in earlier research work [54] , which proved that GMM is an effective representation of distribution load data for a range of distribution load profile types. Finally, plotting the probability density functions for each of the features, revealed that the data follow a multi-modal distribution, which enhanced the argument for using GMM. The idea behind the selection of the dimensionality reduction methods presented earlier in this chapter was to compare the performance of a well established method, such as PCA, and a relatively new one, such as t-SNE. While PCA has been widely used with power systems data, this is not the case with t-SNE, which has been recently gaining interest in other areas and it was decided to explore its effectiveness with power systems data as well. Finally, a selection of classification methods, with different underlying decision surfaces were compared in order to identify those that performed better for the applications discussed in this thesis. The methodology outlined in this chapter was applied on real data in the case studies that are presented and discussed in the next chapters of this thesis. The prediction results for the case studies of Chapters 6 and 7 were assessed using the metrics discussed in the next and final section of this chapter.

4.5 Assessment of Results

In the field of data analysis, there exist a number of performance metrics which are used to assess the results produced by the application of a machine learning method to certain data but not all of them are suitable for all applications.

Sometimes, the classification accuracy of a given method can be sufficient to assess that method's performance for a certain application. However, this metric alone might not always tell the whole story, as there are cases where the classifier might perform poorly even though the classification accuracy can be high. This is explained with the following example regarding a dataset, the examples of which take one of two classes and the majority of these examples belong to one class. It is possible that a classification method trained on a dataset like this, can learn to classify the majority of given examples as if they belonged to the highly populated class. In that case, when some new examples are passed to the learned model, where only few of them (10% for example) belong to one class, the classification accuracy can still be high (in this case 90%), even if all of the passed examples are assigned to the highly populated class. This is why accuracy, precision and recall were selected to assess the results presented later in this thesis, as they can better reflect the quality of predictions.

The expression and meaning for each of the three performance metrics that are used to assess the performance of the classification methods presented throughout this thesis are given below, where:

- True Positive (TP) is a positive example that has been correctly identified as positive,
- True Negative (TN) is a negative example that has been correctly identified as negative,
- False Positive (FP) is a negative example that has been falsely classified as positive, and
- False Negative (FN) is a positive example that has been falsely classified as negative.

The classification accuracy is defined as the ratio of correctly classified examples (both positive and negative) over the total number of examples considered and its expression is shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.29}$$

Precision is calculated by dividing the number of correctly classified positive examples by the total number of examples that were classified as positive, while recall is the number of correctly classified positive examples divided by the total number of positive examples.

$$Precision = \frac{TP}{TP + FP} \tag{4.30}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.31}$$

In the context of fault and event day prediction, which are the subjects of Chapters 6 and 7 respectively, precision refers to the ability of the classifier not to label a *No Fault* (*No Event Day*) example as a *Fault* (or *Event Day*), while the recall refers to its ability to find all *Fault* (*Event Day*) examples.

4.6 Conclusion

This chapter presented the proposed data analysis methodology, which consists of three stages: Characterising Network Behaviour, Anomaly Detection and Prediction of Faults or Disturbances. After a brief overview of the methodology, the purpose of each of the three stages and the relevant techniques that are used throughout the thesis were presented, with more emphasis given on the methods that showed the best performance with respect to the work presented in this thesis. Finally, the assessment criteria, i.e. the metrics that were used to assess the performance of the machine learning methods, were discussed. The work presented in the next chapters involves analysis that corresponds to part or the whole methodology, as this was presented in this chapter. The short case studies of Chapter 5 are brief examples of the "Characterising Network Behaviour" and "Anomaly Detection" stages of the methodology, while the work of

Chapter 6 is a more detailed example of the "Prediction of Faults or Disturbances" stage. A data analysis example covering all three stages of the proposed methodology is given in Chapter 7.

Chapter 5

Identifying Unusual Operation in Distribution Networks: Short Case Studies

This chapter presents some short case studies, before going into the main contributions of this thesis, which are presented in Chapters 6 and 7. These short case studies serve as examples of how data analysis and machine learning can be used to identify unusual operation from distribution network data, when this data is analysed on its own or in conjunction with other data sources, such as weather. While the short case studies presented in this chapter do not involve any kind of fault prediction as the next two chapters, it is worth including these here as they are examples of the first two stages of the proposed methodology and show how unusual operation of the network, possibly unknown to the DNOs, can be identified within existing data. Such an approach could reveal unusual operation relevant to a fault occurrence, which could then form a fault prediction problem.

5.1 Case Study 1: Solar PV Operation

This case study can be seen as an example of the "Anomaly detection" stage of the methodology, as it describes a way of identifying an unusual operating condition, which in this case is the PV operation.

The network behaviour of 6 monitoring locations was compared in the first stage of the analysis, where histograms of the three phase currents were used in order to explore the distribution of the current values for one month (July 2014). It was found, that in 2 out of 6 locations, the current of one phase showed very different behaviour than that of the other two phases. This is illustrated in Figure 5.1, where the 1-min average currents drawn by each phase are shown for each of the monitoring locations.

The axes x and y show the current drawn in amperes (A) by each phase and the number of observations with a specific x value respectively. The L1, L2 and L3 shown in the figure refer to the three phases [1].

Then, the corresponding box plots were produced to investigate the range of current values (1-min average current per phase) based on the hour of the day and, possibly, identify the time of day when the observed behaviour occurred. The current of the phase with the different behaviour was lower during the day, at the times where the solar radiation was expected to be higher. This was an indication that the solar radiation could be related with the observed behaviour in the 2 monitoring locations. As the initial thoughts were to examine any observations with respect to faults and there were no LV faults recorded for any of the monitoring locations during July 2014, it was decided to repeat this process and continue the analysis for one of these locations for July 2015. The selected substation, which will be referred to as NPG Case Study Substation 1, corresponds to the monitoring location 3 from Figure 5.1 and was one of the two monitoring locations, where different network behaviour was observed and where 6 LV faults occurred in July 2015.

First, the behaviour of the network was examined using the load currents to see if it

[1] The convention L1, L2 and L3 is one of the different ways used to refer to the three phases (as is A, B and C).

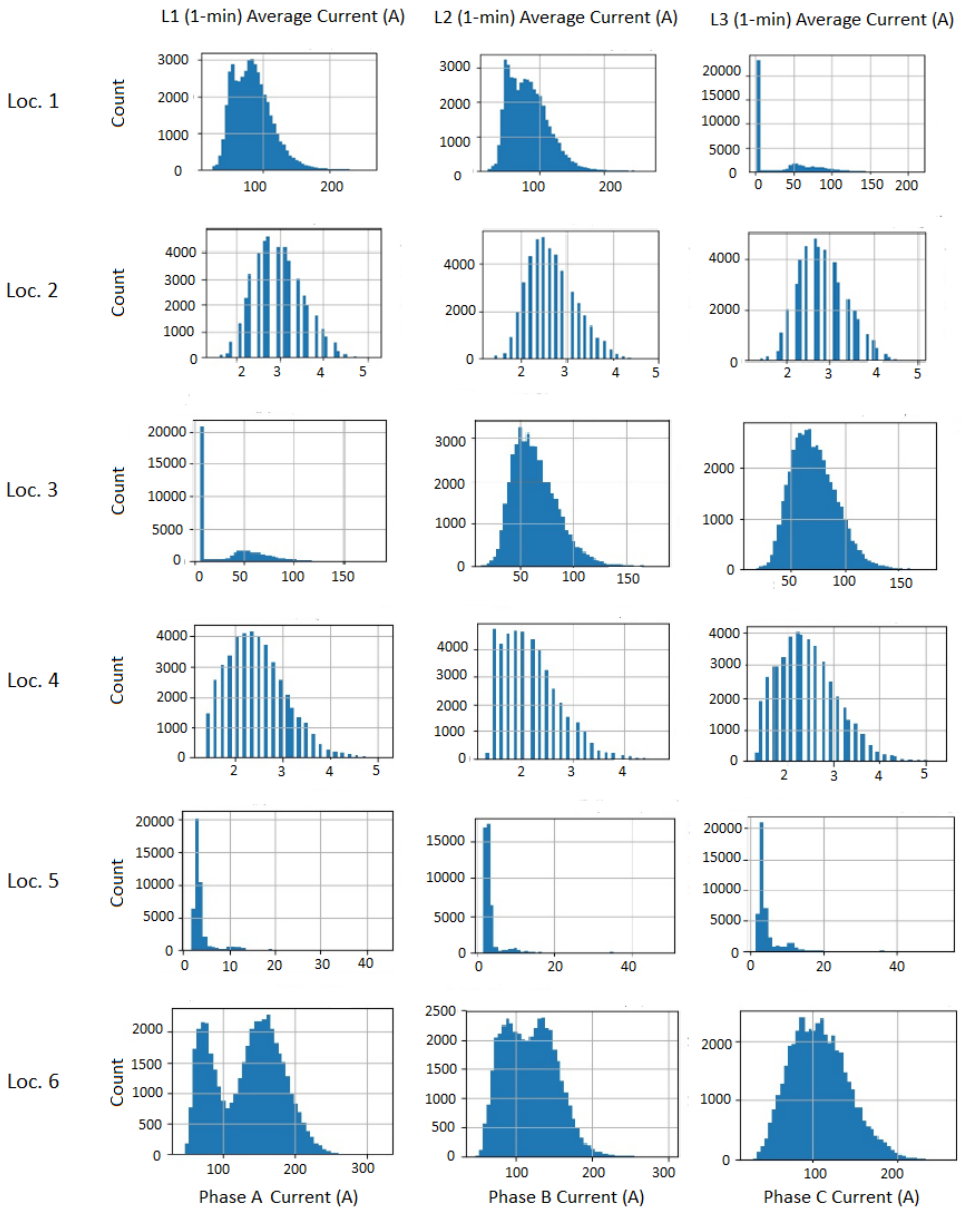


Figure 5.1: Histograms of the 3 phase currents at the 6 monitoring locations for July 2014.

was similar to what we expected from the analysis of the July 2014 data. The initial observations are presented in Figure 5.2, which shows the 1-min average current drawn by the 3 phases at the selected substation.

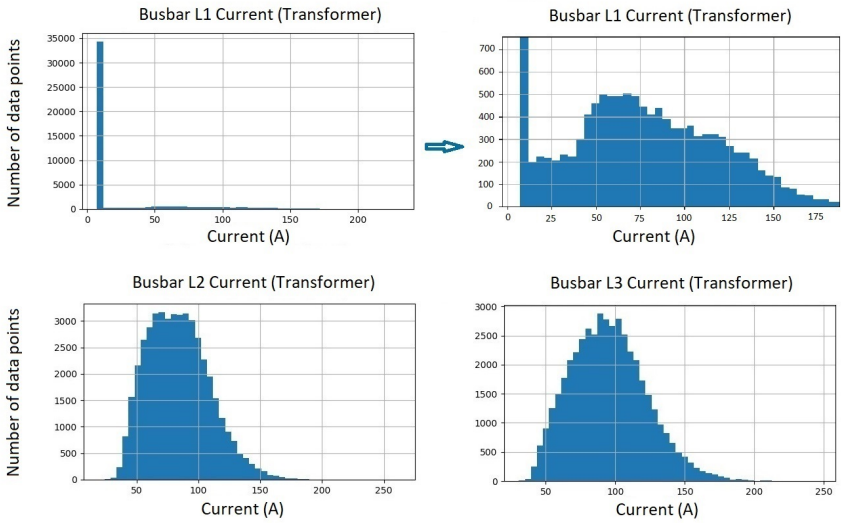


Figure 5.2: Histograms of the 3 phase currents at NPG case study substation 1 in July 2015. The distribution of current values for L1, L2, and L3 are shown in top left, bottom left and bottom right respectively. The top right image is the magnified L1 histogram and shows the distribution of the higher current values.

It can be seen from the histograms in Figure 5.2, that L1 (top left and magnified in top right) shows a very different behaviour than L2 (bottom left) and L3 (bottom right) at this substation. For L1, a very high number of data points take values around 10 A, while the remaining data points follow a distribution which is closer to the distributions of all data points of L2 and L3. The observed current behaviour in July 2015 follows the pattern observed in that same substation in July 2014, which indicates that this pattern corresponds to a specific type of network operation.

Figure 5.3 shows the range of the current values for the 3 phases, based on the hour of the day recorded at this substation for July 2015. It can be seen from the box plots that for L1, the current drops during the day. The median, which is the green line inside the boxes, is $\sim 10-15$ A for all hours and between $09:00-19:00$ the whole box is around this value, which means that 75% of the data points are below $10-15$ A

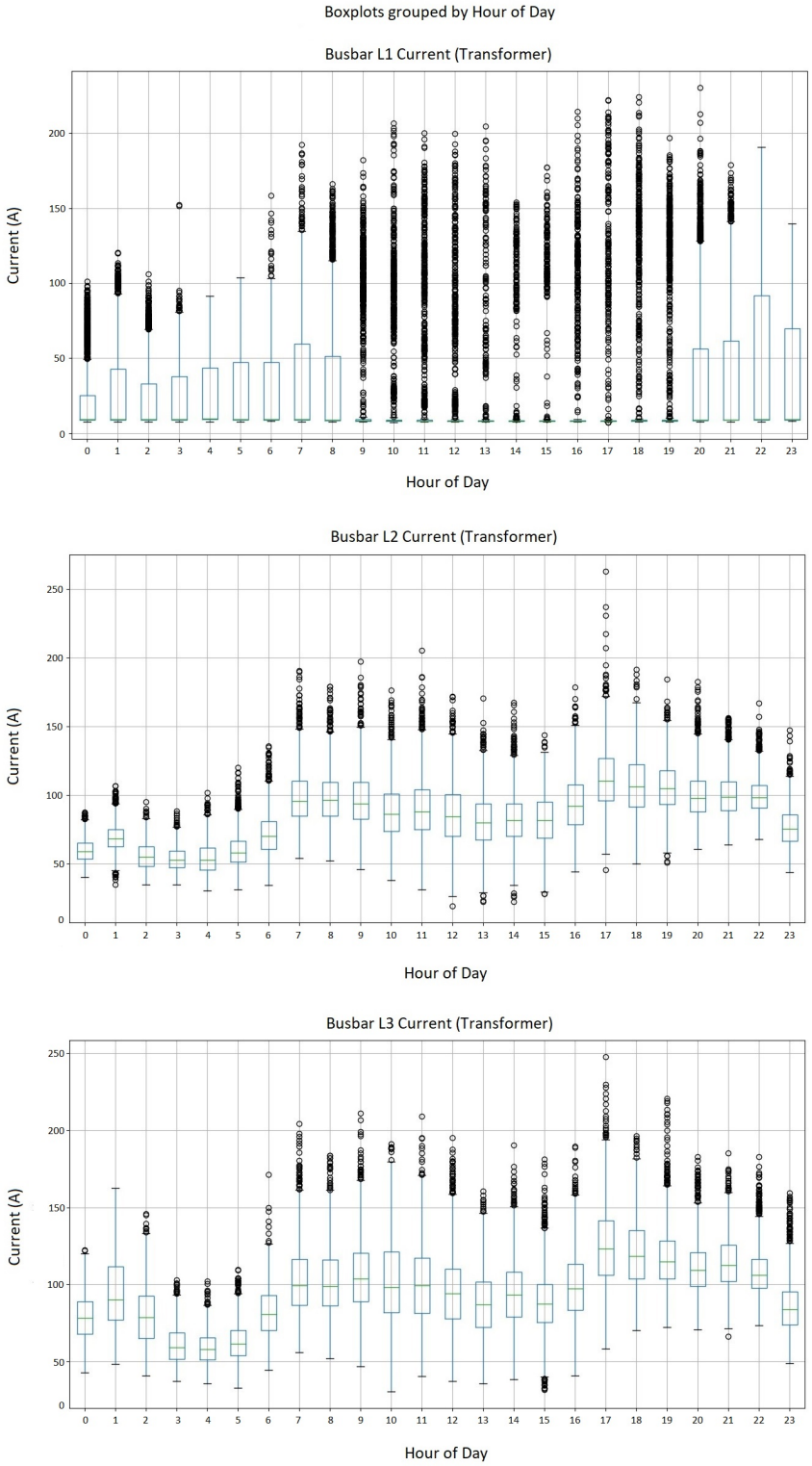


Figure 5.3: Box plots of the 3 phase currents at NPG case study substation 1 in July 2015.

during this time. The L2 and L3 currents are very different than L1 and their values never drop so low.

The operation of PV generation connected to the phase with the reduced current was considered as a possible cause for this difference between the phases.

As PV operation was one of the possible reasons considered for the observed behaviour of L1 current at the NPG Case Study Substation 1, the global radiation observations measured at the nearest Met Office weather station were examined. The boxplot of Figure 5.4, shows the range of values that global solar irradiance takes in July 2015, based on the hour of the day. As expected for a summer month, the solar irradiance measured increases from the early morning hours until the evening hours with a peak at noon. Generally, these are the hours were the L1 load current drops but the extent to which the PV generation contributes to the load current behaviour needs to be further examined.

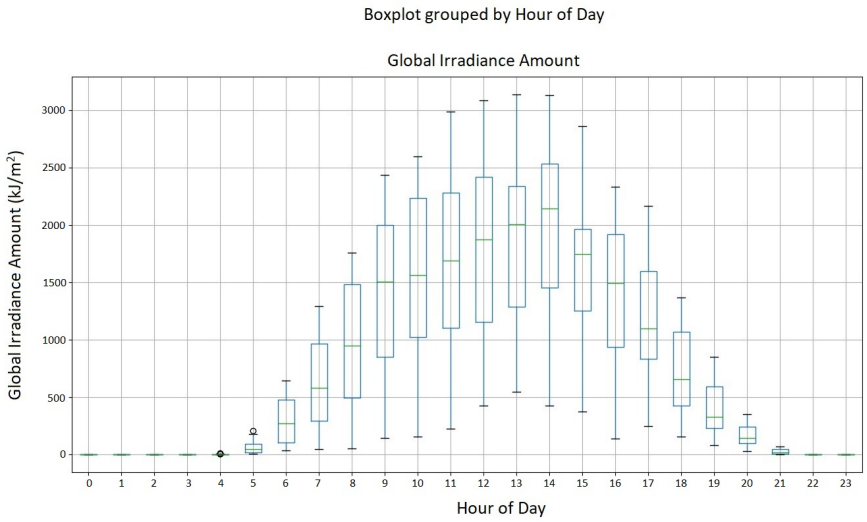


Figure 5.4: Box plot of global irradiance amount measured at the nearest weather station in July 2015.

DNOs could be significantly benefited from the identification of PV operation from power quality data if unmetered, and possibly unknown to them, PV generation is detected, as the uncontrolled widespread PV generation at the LV level could raise voltage and power quality issues. In addition, wide adoption of solar PV could lead

to cables reaching their maximum capacity. Being aware of the amount of PV that is connected to their network would allow DNOs to anticipate such developments and plan for investment when necessary. This would lead to a better understanding of their distribution network and give them the opportunity to develop a more efficient way of managing their network.

5.2 Case Study 2: Phase Imbalance

The work presented in this case study can be considered as another "Anomaly detection" example, as it shows how a ternary plot can be used to identify a case of phase imbalance. Then, some further analysis is conducted, in which the relation between unbalance and temperature is examined.

In Section 5.1 above, the load currents of the three phases were examined and it was found that the load current of Phase A was generally lower than the currents of Phases B and C. As the total load current is divided between the three phases we can think of the current measurements as compositional data, which can be represented by the percentages of the current of each phase with respect to the total load. Compositional data carries only relative information, as they represent parts of some whole and can be represented by a ternary plot [90]. The ternary plot used for the representation of the phase current percentages in our case can be seen in Figure 5.5.

Each of the data points shown in Figure 5.5 corresponds to a 1 - minute measurement and its position in the ternary plot is determined by the distribution of the load across the three phases for that minute. If the load was equally distributed among the three phases at any point in July, the data points would be mainly concentrated in the centre of the triangle. Figure 5.5, however, implies that there are times where the load at NPG Case Study Substation 1 is balanced (as there is a large amount of data points concentrated in the centre of the triangle) and times where the load supplied by Phases B and C is much higher than that of Phase A (as there is a large group of data points concentrated in the left side of the triangle). This representation of the data is consistent with the observations made in Section 5.1 regarding the load currents.

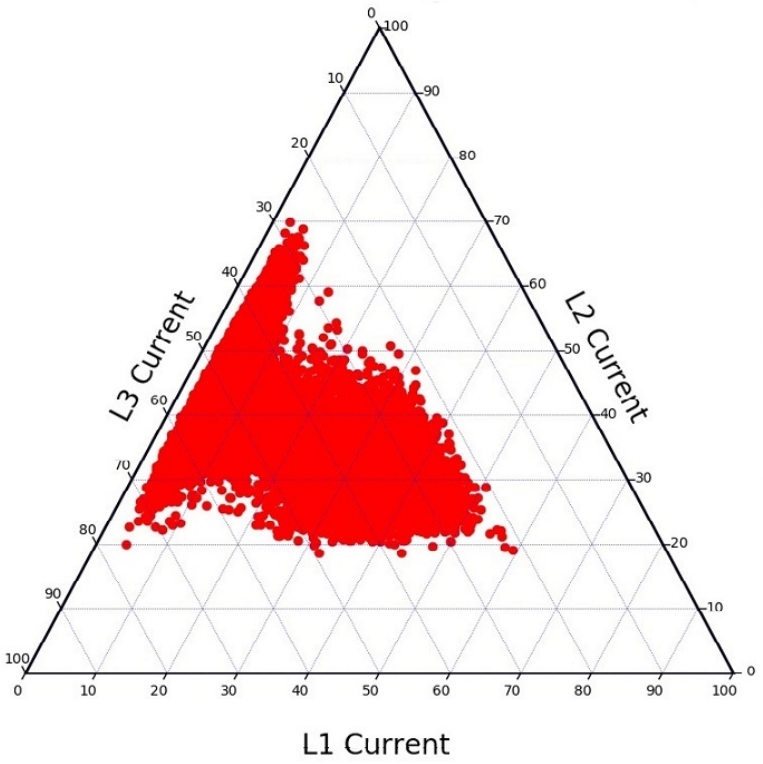


Figure 5.5: Ternary plot for the phase current percentages in July 2015.

One could think that the data points concentrated on the left side of the triangle would correspond mainly to the daylight hours, as the possible PV operation that was identified in Section 5.1 is likely to have a contribution to the observed unbalance. However, this was not the case when the ternary plot was examined based on the hour of day, as for every hour of the day, there were data points on both groups as well as between them.

In order to examine another possible factor that could contribute to the observed behaviour, the relationship between the measured unbalance and the temperature was investigated. It is worth noting that the measured unbalance that is considered in this case study was a separate measurement recorded by the monitoring device and was not derived from the current measurements which were used for the ternary plot of Figure 5.5. It can be seen from Figure 5.6, that the unbalance and the temperature vary in a similar way in July 2015.

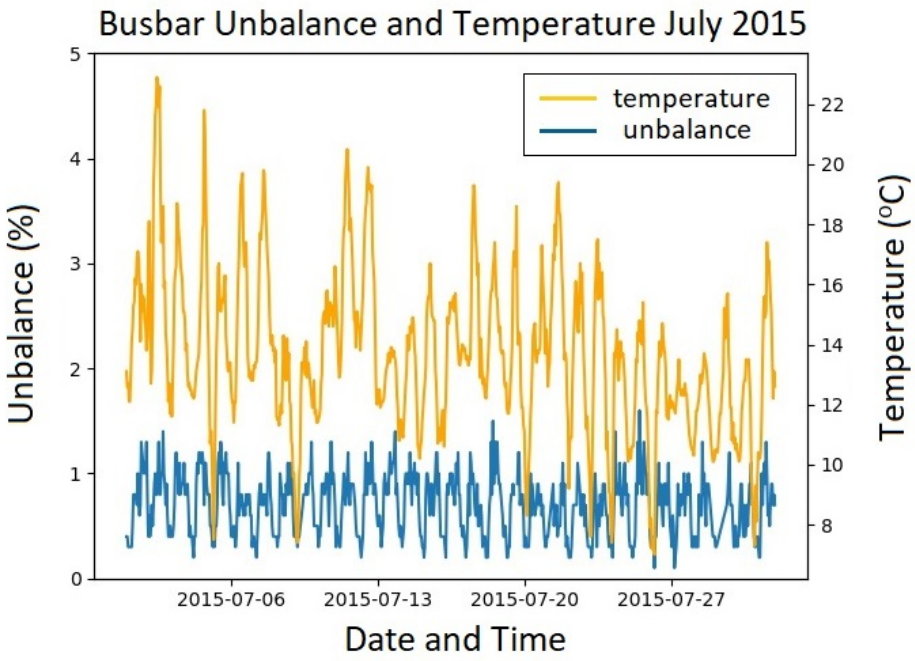


Figure 5.6: Hourly unbalance vs temperature in July 2015.

Quantile regression was used to examine the impact of temperature on different quantiles of the unbalance $(5\%, 15\%, 25\%, ..., 95\%)$ and the result can be seen in Figure 5.7. The different quantiles are represented by the grey lines, while the red line corresponds to the ordinary least squares model. A first observation that could be made is that the mode of unbalance shifts upwards as the temperature increases, as indicated by the positive slope. Moreover this trend is clearer for low or high temperatures (but less so for moderate temperatures), which also indicates that the variance does not remain constant as the temperature changes. This is the reason that quantile regression, rather than ordinary least squares, was used to extract more relevant information. In addition, Figure 5.7 shows that the ordinary least squares model misses the median imbalance at the low temperatures but gets closer at the higher ones. This means that quantile regression is more suitable to capture the extreme behaviour, which would be of interest here, unlike a typical regression model such as the ordinary least squares which would generate false alarms at lower temperatures.

Further work is required in order to establish causal relations between unbalance and

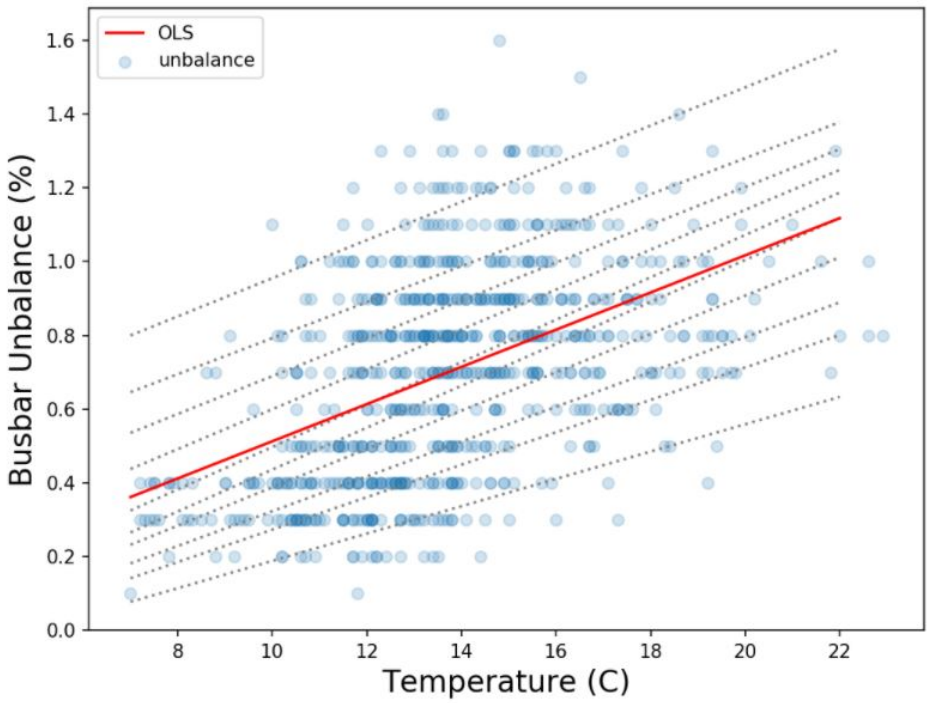


Figure 5.7: Quantile regression.

weather conditions or the presence of distributed generation or unbalanced load in the distribution network.

Being able to establish such relations would be beneficial to the DNOs as it could help them mitigate impacts, such as losses, that unbalanced load has on the distribution network.

5.3 Case Study 3: Unusual Network Behaviour Detection Using Dimensionality Reduction

The primary aim of this case study was to demonstrate some of the benefits of data analysis and how it could be used towards fault anticipation in order to improve network operation. To do this, data science techniques were applied on real network data to identify patterns of unusual behaviour. The data analysed in this case study was gathered from ENW's distribution network during the Low Carbon Network Fund

(LCNF) project Capacity to Customers (C2C). Power quality monitoring devices, called PQubes [91], were used to gather the data. These devices measure quantities such as phase currents and voltages, harmonic currents and voltages, real and reactive power, frequency, flicker, Total Harmonic Distortion (THD) etc, and have the ability to record a number of power quality disturbances.

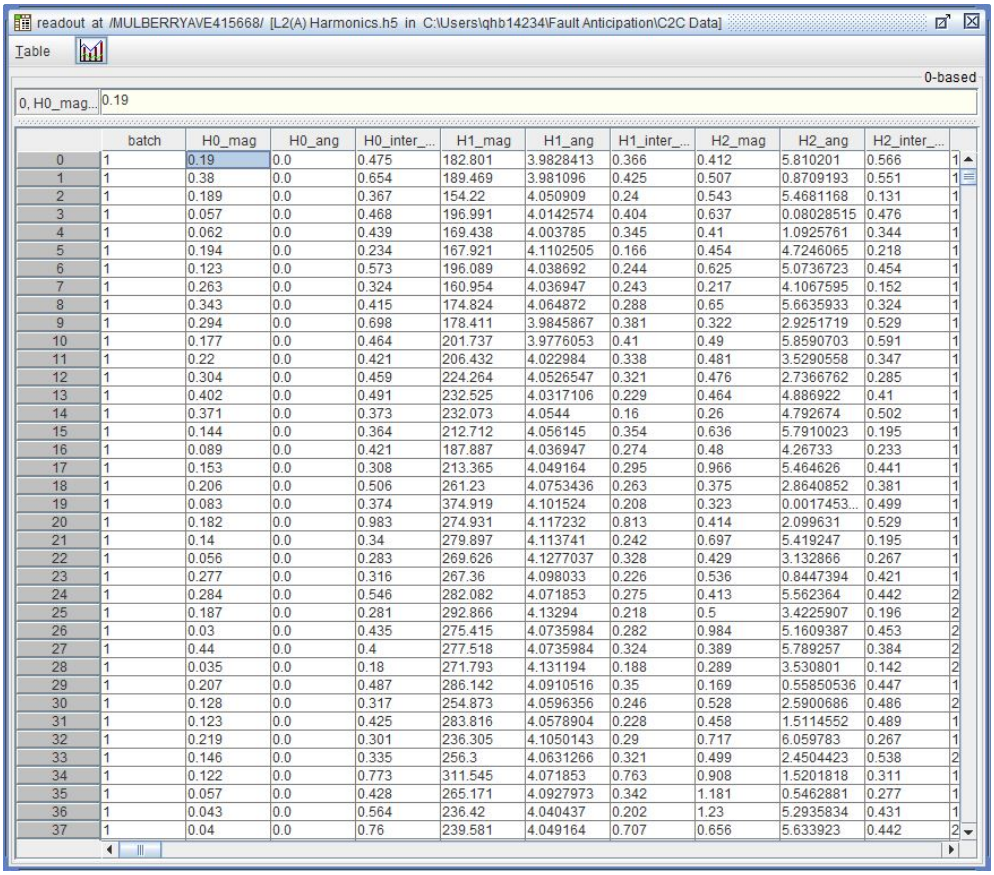
For this case study, only the harmonic datasets were analysed. The data consists of six datasets, one for the voltage (in Volts) and one for the current (in Amperes) of each of the three phases. Each dataset contains measurements for the DC component (mentioned in the datasets as 0th harmonic), the fundamental frequency and harmonics (up to the 63rd harmonic) for a total of 77 substations. For the analysis described in this report, only data up to (and including) the 7th harmonic was considered, for 12 out of 77 substations. The reason for this selection was that the third, fifth and seventh harmonics are the most common and account for much of the distortion of the current signal, while higher order harmonics have generally less impact [92]. For each of the harmonics, the monitored quantities are the magnitude, the angle and the interharmonic magnitude and the measurements are averaged in 15 minute intervals. The substation measurements were, generally, gathered from end of February 2013 to June 2014 but the specific dates of available data vary for the different substations. Indicatively, to show the form in which the harmonic datasets were available, part of the data for the phase B current of one of the substations is shown in Figure 5.8.

The characteristics of the data used for the analysis are summarised in Table 5.1.

Table 5.1: Summary of C_2C Data Used for the Analysis

Quantities Analysed	I(Amps), V(Volts) - fundamental, dc and harmonic data up to the 7^{th} harmonic (magnitudes and angles) and interharmonic magnitudes
Data Resolution	15-min averaged data
Voltage Level	Measurements at the LV side of the 11/0.4 kV transformers
Duration	Spring 2013 - Summer 2014

Before analysing the data, the datasets were unified so that the analysis for each substation would look into all voltage and current measurements at the same time. This resulted in the unified data having 144 dimensions (6 datasets, with 24 columns



readout at /MULBERRYAVE415668/ [L2(A) Harmonics.h5 in C:\Users\qhb14234\Fault Anticipation\C2C Data]

Table

0-based

0, H0_mag... 0.19

	batch	H0_mag	H0_ang	H0_inter_...	H1_mag	H1_ang	H1_inter_...	H2_mag	H2_ang	H2_inter_...	
0	1	0.19	0.0	0.475	182.801	3.9828413	0.366	0.412	5.810201	0.566	1
1	1	0.38	0.0	0.654	189.469	3.981096	0.425	0.507	0.8709193	0.551	1
2	1	0.189	0.0	0.367	154.22	4.050909	0.24	0.543	5.4681168	0.131	1
3	1	0.057	0.0	0.468	196.991	4.0142574	0.404	0.637	0.08028515	0.476	1
4	1	0.062	0.0	0.439	169.438	4.003785	0.345	0.41	1.0925761	0.344	1
5	1	0.194	0.0	0.234	167.921	4.1102505	0.166	0.454	4.7246065	0.218	1
6	1	0.123	0.0	0.573	196.089	4.038692	0.244	0.625	5.0736723	0.454	1
7	1	0.263	0.0	0.324	160.954	4.036947	0.243	0.217	4.1067595	0.152	1
8	1	0.343	0.0	0.415	174.824	4.064872	0.288	0.65	5.6635933	0.324	1
9	1	0.294	0.0	0.698	178.411	3.9845867	0.381	0.322	2.9251719	0.529	1
10	1	0.177	0.0	0.464	201.737	3.9776053	0.41	0.49	5.8590703	0.591	1
11	1	0.22	0.0	0.421	206.432	4.022984	0.338	0.481	3.5290558	0.347	1
12	1	0.304	0.0	0.459	224.264	4.0526547	0.321	0.476	2.7366762	0.285	1
13	1	0.402	0.0	0.491	232.525	4.0317106	0.229	0.464	4.886922	0.41	1
14	1	0.371	0.0	0.373	232.073	4.0544	0.16	0.26	4.792674	0.502	1
15	1	0.144	0.0	0.364	212.712	4.056145	0.354	0.636	5.7910023	0.195	1
16	1	0.089	0.0	0.421	187.887	4.036947	0.274	0.48	4.26733	0.233	1
17	1	0.153	0.0	0.308	213.365	4.049164	0.295	0.966	5.464626	0.441	1
18	1	0.206	0.0	0.506	261.23	4.0753436	0.263	0.375	2.8640852	0.381	1
19	1	0.083	0.0	0.374	374.919	4.101524	0.208	0.323	0.0017453...	0.499	1
20	1	0.182	0.0	0.983	274.931	4.117232	0.813	0.414	2.099631	0.529	1
21	1	0.14	0.0	0.34	279.897	4.113741	0.242	0.697	5.419247	0.195	1
22	1	0.056	0.0	0.283	269.626	4.1277037	0.328	0.429	3.132866	0.267	1
23	1	0.277	0.0	0.316	267.36	4.098033	0.226	0.536	0.8447394	0.421	1
24	1	0.284	0.0	0.546	282.082	4.071853	0.275	0.413	5.562364	0.442	2
25	1	0.187	0.0	0.281	292.866	4.13294	0.218	0.5	3.4225907	0.196	2
26	1	0.03	0.0	0.435	275.415	4.0735984	0.282	0.984	5.1609387	0.453	2
27	1	0.44	0.0	0.4	277.518	4.0735984	0.324	0.389	5.789257	0.384	2
28	1	0.035	0.0	0.18	271.793	4.131194	0.188	0.289	3.530801	0.142	2
29	1	0.207	0.0	0.487	286.142	4.0910516	0.35	0.169	0.55850536	0.447	1
30	1	0.128	0.0	0.317	254.873	4.0596356	0.246	0.528	2.5900686	0.486	2
31	1	0.123	0.0	0.425	283.816	4.0578904	0.228	0.458	1.5114552	0.489	1
32	1	0.219	0.0	0.301	236.305	4.1050143	0.29	0.717	6.059783	0.267	1
33	1	0.146	0.0	0.335	256.3	4.0631266	0.321	0.499	2.4504423	0.538	2
34	1	0.122	0.0	0.773	311.545	4.071853	0.763	0.908	1.5201818	0.311	1
35	1	0.057	0.0	0.428	265.171	4.0927973	0.342	1.181	0.5462881	0.277	1
36	1	0.043	0.0	0.564	236.42	4.040437	0.202	1.23	5.2935834	0.431	1
37	1	0.04	0.0	0.76	239.581	4.049164	0.707	0.656	5.633923	0.442	2

Figure 5.8: Phase B current data for one of the 77 C_2C substations.

each).

First, the process that was followed during the analysis is described and then the results are presented with graphs and discussed later in a subsection of this section.

Initially, the t-SNE technique was applied to the data of a number of different substations, in order to visualise the high dimensional data and investigate the relation of the data points with each other. Due to the duration of the available not being the same, the number of data points for each substation was different. The data used for analysis for Whitefield Rd substation (which is the focus of this section), had 144 dimensions and 44736 data points. However, for computational purposes, only 10000 randomly selected data points were fed into t-SNE. It was found that the behaviour of two substations (Whitefield Rd and Mulberry Ave) was different than that of the other

substations that were examined. The differences in the observed substation behaviour are explained in more detail later.

The next steps of the analysis focused on these two substations, due to the observed difference in their behaviour. As the possible relation of any usual observation with the occurrence of faults was of interest to this project, more emphasis was given on Whitefield substation as there was a known fault recorded on this substation in the C2C reports [93]. Whitefield Rd and Mulberry Ave substations are 11kV/0.4kV and both are supplied from the same primary substation (Holme Rd substation 33kV/11kV).

In order to explain the different behaviour, the data were examined based on the different seasons to see if there was a seasonal variation in these two substations that was not present in the others. It was found that the two substations behaved differently throughout the year, something that was not observed in the other substations examined. The different behaviour was mostly during spring, and because the data contained measurements for both spring 2013 and 2014, it was decided to treat spring 2013 and spring 2014 as different seasons for the rest of the analysis.

The seasons were defined as:

- Spring 2013: 1 March 2013 — 31 May 2013
- Summer: 1 June 2013 — 31 August 2013 and few days of June 2014
- Autumn: 1 September 2013 — 30 November 2013
- Winter: 1 December 2013 — 28 February 2014 and few days of February 2013
- Spring 2014: 1 March 2014 — 31 May 2014

In order to examine what caused the seasonal variation that was observed, the data was examined again in different ways (e.g. looking at each harmonic order separately or at voltages and currents separately). The results of the data analysis process that was described here are presented below.

5.3.1 Main Observations

This subsection is divided into four parts which describe and discuss the main observations. The visualisation of the high dimensional data in the two dimensional space using t-SNE will be shown in this section using scatter plots. The same amount of measurements (usually 2000 points) were randomly chosen from each season and fed into t-SNE, in order to examine the behaviour of each substation throughout the year. The scatter plots are coloured based on the different seasons (defined earlier) and the colours that correspond to each season are shown in Table 5.2:

Table 5.2: Season Colours in Scatter Plots

Red	Spring 2013
Green	Summer
White	Autumn
Blue	Winter
Yellow	Spring 2014

Observation 1: Whitefield Rd and Mulberry Ave Substations

For the majority of the substations examined, t-SNE produced a scatter plot similar to the one shown in Figure 5.9.

In the scatter plot of Figure 5.9, no seasonal variation can be observed, as data points from all seasons show no spatial segregation. The scatter plot shown in Figure 5.9 corresponds to the Ramsden Special School secondary substation.

Unlike the rest of the substations, Whitefield Rd and Mulberry Ave show a very different behaviour as it can be seen in Figures 5.10 and 5.11 respectively.

In Figure 5.10, it is evident that in Whitefield Rd substation spring 2013 (red) shows a very different behaviour than the rest of the year, as almost all data points belonging to that spring are gathered in a cluster that is separate from the rest of the points.

The scatter plot in Figure 5.11 that corresponds to the Mulberry Ave substation shows an even more diverse behaviour than that of Whitefield substation, as not only spring 2013 (red) but also summer (green) data points are separate from the points that

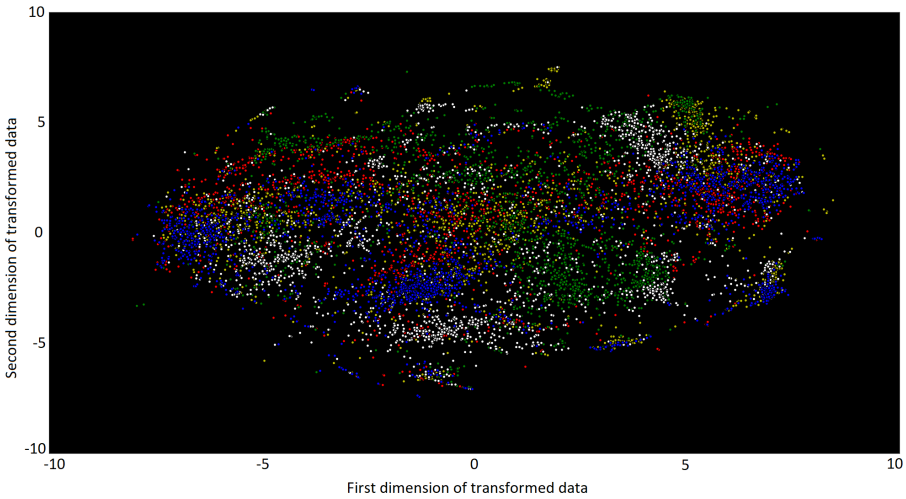


Figure 5.9: t-SNE visualisation on data for Ramsden Special School substation. The different colours of the data points represent the seasons they belong to as defined in Table 5.2.

belong to the other seasons. In both Whitefield and Mulberry substations, the general behaviour of autumn, winter and spring 2014 is similar and resembles that of the other substations.

Specific observations about Whitefield substation, where the analysis focused are discussed below.

Observation 2: Week 4-10 April 2013

In Figure 5.10, apart from the red cluster where most of the spring 2013 points are gathered, there is a "line" of red points that are separated from the data points belonging to both the spring 2013 and the rest of the year. These data points are shown in the red circle in Figure 5.12. A similar group of spring 2013 data points was present in all the different scatter plots produced by the t-SNE for different random points selected each time. When examined, these data points were found to belong in the week 4-10 April 2013. After looking again into the raw data, it was found that during this week in April, the current drawn by all 3 phases is almost doubled. This can be seen in Figure 5.13, which shows the phase A current at Whitefield substation. It can be seen

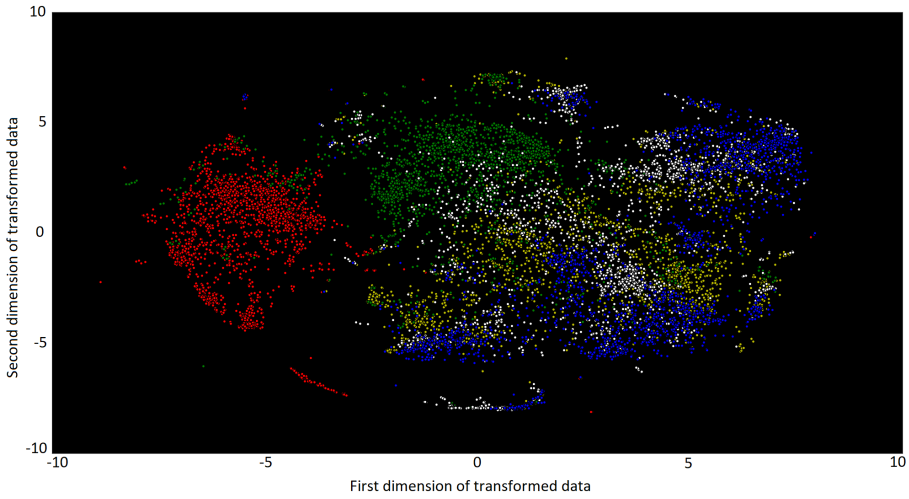


Figure 5.10: t-SNE visualisation on data for Whitefield Rd substation. The different colours of the data points represent the seasons they belong to as defined in Table 5.2.

that the current starts increasing on the 4^{th} of April and decreases again on the 10^{th}.

A possible reason for this observation could be an ongoing maintenance in an adjacent feeder and load transferred to the monitored feeder.

Observation 3: Spring 2013

It has already been mentioned that, in Whitefield substation, the spring of 2013 shows a very different behaviour than the rest of the year and the spring of 2014. This observation implies that something unusual happened in that spring in the area of Whitefield substation. After an online search regarding the weather of spring 2013, it was found that it was the "coldest spring for the UK since 1962 (marginally colder than spring 1979), and the fifth coldest in a series since 1910" [94]. Obviously, this could not be the only reason why this behaviour is observed as the weather would affect other locations as well. However, it could be an indicator that there are specific elements in the network, the operation of which is affected by the weather.

In order to further investigate this unusual behaviour, the DC component, the funda-

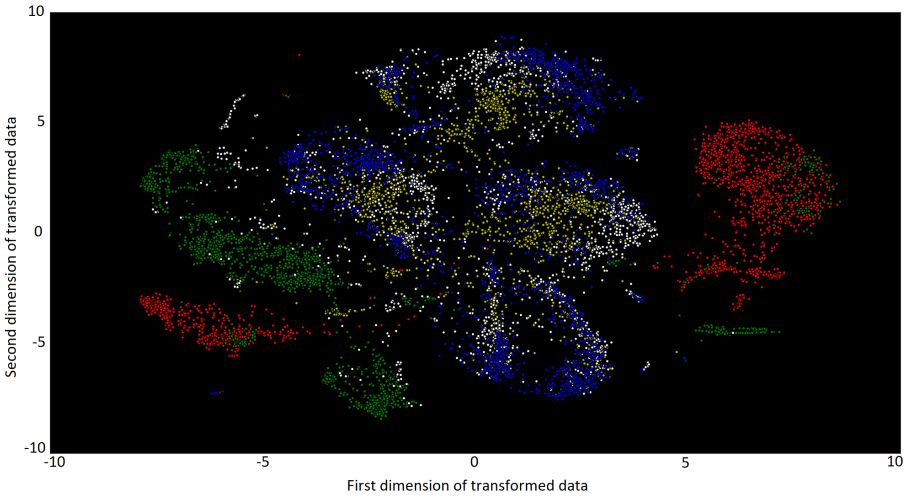


Figure 5.11: t-SNE visualisation on data for Mulberry Ave substation. The different colours of the data points represent the seasons they belong to as defined in Table 5.2.

mental frequency and the different harmonics were examined separately. An interesting finding was that the DC component manifests a similar behaviour to that observed in the Whitefield substation when all the data was considered. The scatter plot for the DC component is shown in Figure 5.14.

In Figure 5.14, where only the DC components of the voltages and currents are considered, the data points corresponding to spring 2013 are clearly separated from the rest. This was not the case for the fundamental frequency or any of the other harmonics. Although the shapes of the scatter plots corresponding to the fundamental and the harmonic frequencies were different to each other, the spring 2013 points were generally mixed with the data points from all other seasons. When the DC component was examined in the raw data, it was found that the DC component of the voltages was indeed higher during spring 2013 compared to the other seasons. This is illustrated in Figure 5.15, where the DC component of the phase A voltage is shown.

The DC component in a power system can be affected by a number of factors, including: (a) Geo-magnetically induced currents (b) PV systems connected without mains frequency transformers (c) Computers and other DC loads (d) AC and DC drives (e) HVDC systems when AC and DC transmission lines are in close proximity (f) Static

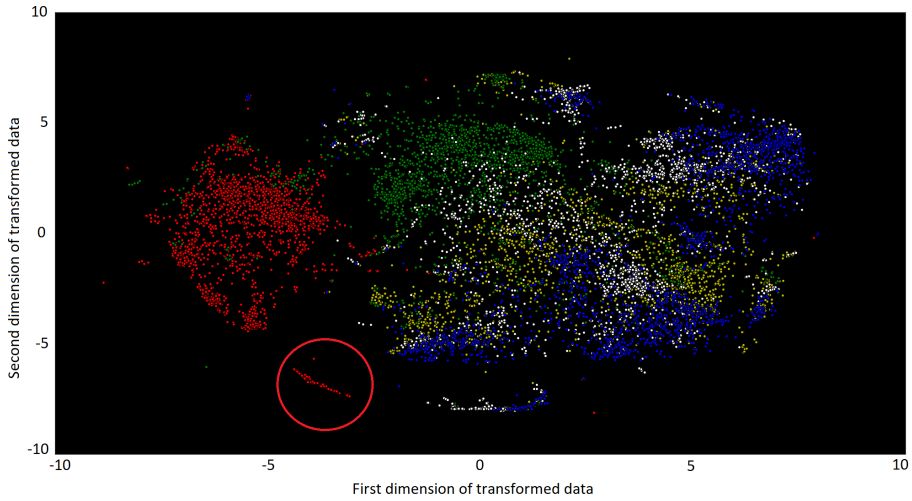


Figure 5.12: Data points corresponding to the week 4-10 April 2013 in Whitefield Rd substation.

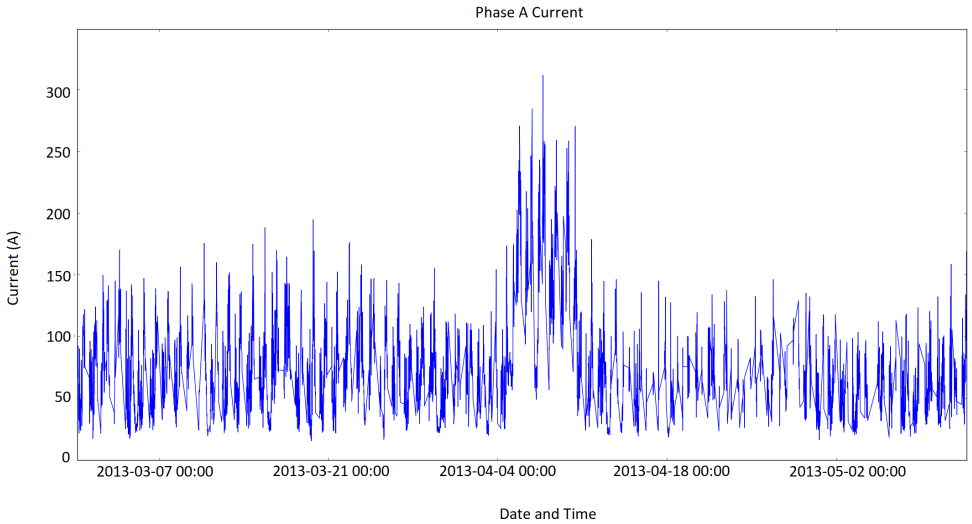


Figure 5.13: Phase A current in Whitefield Rd substation.

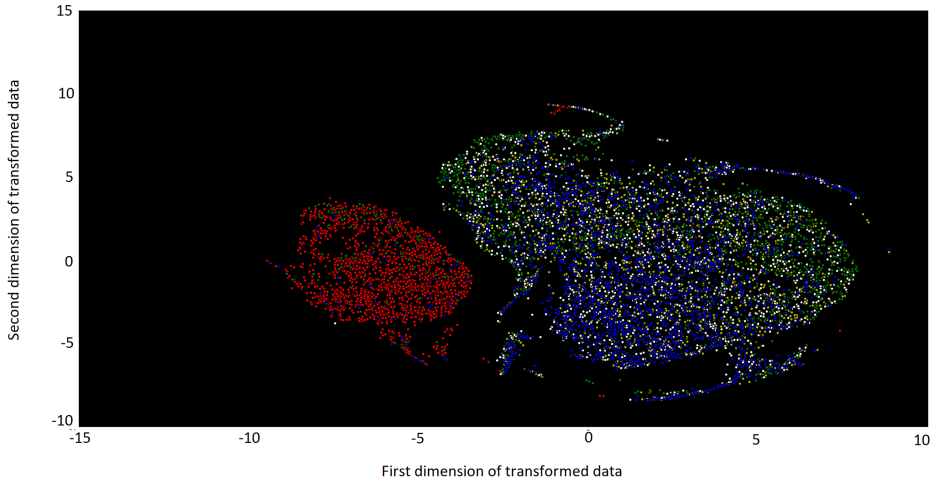


Figure 5.14: t-SNE on DC component data only for Whitefield Rd substation.

VAr compensators.

Combining the fact that the spring of 2013 was the coldest spring in 51 years with the higher DC component measured in the voltages at Whitefield substation, it can be deduced that there might be a relation of the connected load (for example electrical heating) in the area of Whitefield substation to the different behaviour of that spring. Unfortunately, there was not enough information or data available that could allow further investigation of any potential relation of the connected load (or even generation) to the unusual behaviour observed in spring 2013. Although the DC component seems to have an impact on the behaviour of spring 2013, it cannot be the only factor affecting it, as the behaviour of spring remains different even when the DC component is removed from the data.

Possible reasons for the observed change in the voltage DC components at this time could be related to monitoring issues, perhaps a change or problem of the monitoring device in the beginning of the summer or an intervention by the DNO that would explain the change in the DC component.

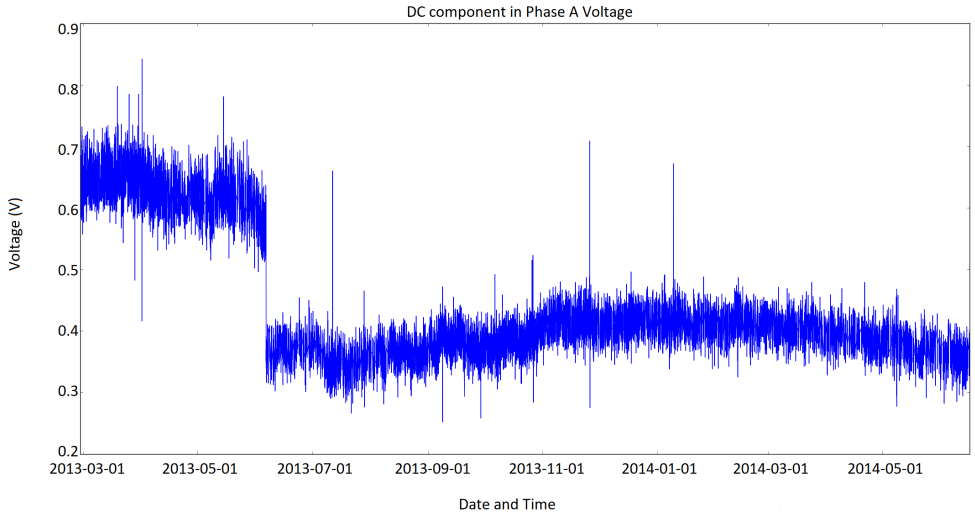


Figure 5.15: DC magnitude of phase A voltage for Whitefield Rd substation.

Observation 4: Autumn/Winter Outliers

Another group of outliers that was present in all the different runs was the "line" consisting of autumn and winter points that is shown in the blue circle in Figure 5.16. With a closer look into the data points, it was found that even though most of the points in this group belong to autumn or winter, there were also some points from summer or spring. When examined, most of the points were found to belong to random dates, seemingly unrelated, but when checked, the specific dates were found to be mostly in the weekend (mostly Sundays) or holidays (Christmas and New Year's day were also found in this group). More analysis is required to find the exact reason that causes these data points to separate from the rest, but the fact that most of the points belong to weekends or holidays in the autumn or winter could be an indication that the observed behaviour could be due to an increase in the load, as people are more likely to stay at home in the colder weekends of autumn or winter, as well as on holidays such as Christmas and New Year's day.

During this case study, it was shown that it is possible to identify cases of unusual operation of the network using t-SNE. Although the analysis was incomplete and there

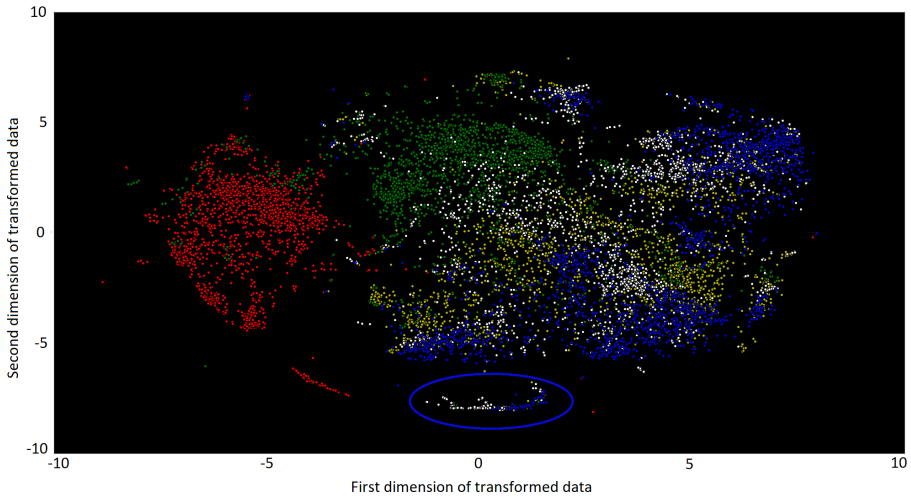


Figure 5.16: Autumn/winter outliers in Whitefield Rd substation.

are no final answers on the exact causes the observed behaviour, a number of significant conclusions can be drawn. Even with 15-minute average data, it was possible to identify unusual behaviour within the data. Spatial and seasonal variation was observed as the behaviour of two neighbouring substations was found to be different than other substations across ENW's distribution network and that behaviour was very different in spring 2013 compared to the rest of the year. It is believed that higher frequency of data could reveal more aspects of the network's behaviour, which combined with additional data such as load and generation types, fault records, historical network data etc, would enable a more in depth analysis of the distribution network and would help identify patterns of pre-failure activity that could be used for fault anticipation.

As the purpose of this case study, which was to demonstrate potential benefits of this type of analysis, was achieved it was decided not to continue the analysis of this data, as this would involve discussions with ENW and request of additional data. It was decided to start analysing the first group of Northern Powergrid data instead, while taking into account the observations made during this case study.

5.4 Conclusion

This chapter presented three short case studies, which aimed to illustrate how the proposed methodology of Chapter 4 can be used to identify unusual operation of the network. By analysing distribution network data, either on its own or in conjunction with weather data, typical behaviour (Characterisation) and cases where the data did not show the expected behaviour (Anomaly Detection) were identified using visualisation. The observed behaviour was further explored in order to identify possible causes. The purpose of the work presented in this chapter was to show how the proposed methodology can be used on existing data and repurpose it to identify potential issues, possibly unknown to the DNOs, rather than providing detailed solutions to the identified problems.

This is done in the next two chapters which demonstrate how the proposed methodology can be used to test and confirm DNO conjectures based on expert knowledge, by translating heuristics to formal models. Building on observations made by DNOs on how different factors seem to affect the occurrence of faults in their networks, the work presented in the next two chapters demonstrate how the proposed methodology can be used to identify relations within the available data and to use these relations in order to predict the occurrence of distribution network faults or disturbances, which is the main topic addressed in this thesis.

Chapter 6

Weather-Related Fault Prediction in Minimally Observed Distribution Networks

The weather related impacts on the power system and the uncertainties accompanying climate change have been a research subject of the academic and power industry sectors. However, as discussed in Chapter 2, the topic of weather-related fault prediction considering only the weather conditions in the area of the fault has not been thoroughly addressed and is the topic of this chapter.

6.1 Overview of Weather Fault Prediction Case Study

This work examines the consequence of weather conditions on the distribution network operation and attempts to predict the occurrence of weather-related faults in the case where only weather observations are available. These are typical conditions for a UK distribution network, where little to no monitoring is usually available. In addition, this work was conducted with a UK DNO and stemmed from expert observations

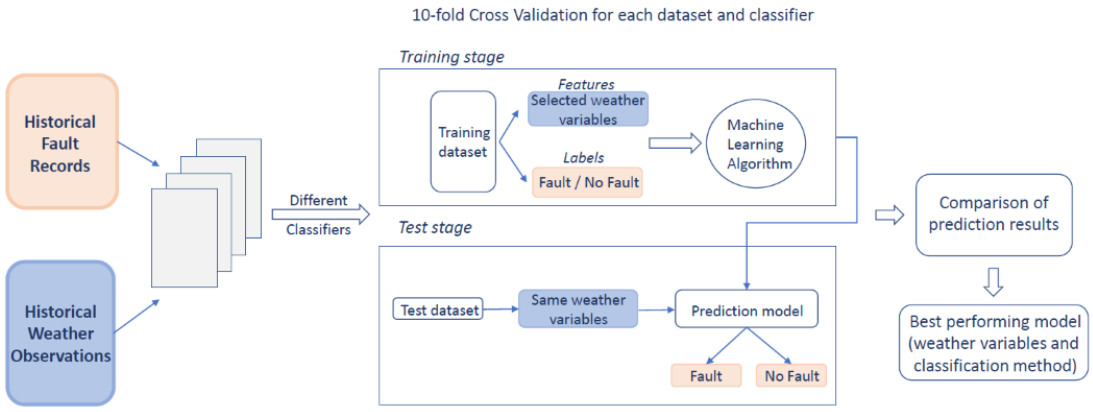


Figure 6.1: Data analysis methodology for weather-related fault prediction.

that certain faults occurred following certain weather conditions, therefore it is an application with real operational value. Therefore, the methodology and predictive models from this research can be deployed immediately to offer value in a distribution network. There is no need for costly additional monitoring equipment installation. This is also achieved without detailed topographic information or widespread environmental sensing. The focus of this chapter is on the prediction of weather related faults at the HV and LV level of a distribution network, considering only the fault history for that network and historical weather conditions. The purpose of this work is not to predict the exact location of the fault or specific type of electrical fault, but the circuit that is more likely to be affected by a fault, given the expected weather conditions in that area. To this end, statistical analysis and machine learning methods are used, where the fault records are used as ground truth for the event occurrences in the network and minimal environmental data are examined as fault causes. The data analysis methodology presented in this chapter can be summarised in Figure 6.1 The work presented in this chapter corresponds to the 'Prediction of Faults or Disturbances' stage of the methodology proposed in Chapter 4 and details the process that was followed in order to determine a suitable predictive model.

The diagram shown in Figure 6.1 gives an overview of how the weather variables are

analysed in conjunction with the fault records in order to identify a suitable fault prediction model. The different stages of this chapter's methodology from the dataset development to the selection of the fault prediction model of Figure 6.1 are detailed later in this chapter. The following sections cover the network operator data and context, the machine learning methods and the data analysis process as well as the results which are presented in the form of three case studies.

6.2 Network Operator Data and Context

The work presented in this chapter utilises fault data from a real distribution network alongside historic weather observations. This section provides an overview of the context and nature of the data used.

6.2.1 Fault Records

Five years of fault records were provided by Northern Powergrid and used for the analysis. Northern Powergrid, which is one of the DNOs in the United Kingdom, is responsible for the distribution networks in the North East, Yorkshire and northern Lincolnshire. The fault record files contained the incidents that occurred at both the HV and LV levels of their distribution network and were examined separately.
The HV fault records cover the period 20/05/2013 – 20/07/2018 and contain 17653 events in total. The range of voltages covered in the HV fault records was 6.6kV - 132kV with the majority of recorder faults occurring at 11kV (76%) followed by 20kV (18%), which is reasonable as the largest part of NPG's distribution network operates in these levels. The postcode of the incident location was available in the report description for 16318 of these faults, with 2441 of them being weather related faults (based on the cause registered in the fault records). The causes included in the weather related faults and the number of events per cause are listed in Table 6.1.

Looking at the distribution of the HV faults throughout the year and their causes, as presented in Figure 6.2, the following observations can be made:

Table 6.1: HV Weather Related Faults

Cause	**No. of Events**
Wind and Gale (excluding Windborne Material)	1023
Lightning	902
Snow, Sleet and Blizzard	204
Windborne Materials	133
Flooding	70
Ice	34
Rain	34
Solar Heat	25
Freezing Fog and Frost	16

- "Lightning" is the most prevalent cause of HV weather-related faults in the summer months, with July being the month with the highest number of faults due to lightning.
- "Wind and gale (excluding windborne material)" is the most prevalent cause of HV weather-related faults in the winter months, with the majority of faults due to wind and gale occurring in December.

It is worth noting here that there is a distinction between "Rain" and "Flooding" faults in the above table, even though the latter occur mainly when there is heavy rain. The cause of faults that occur as a result of damage to overground or underground electrical equipment following a flood in certain areas is considered to be "Flooding", while "Rain" faults include faults due to rainwater permeating into damaged cables or equipment.

The locations of the fault occurrences on Northern Powergrid's HV network can be seen in Figure 6.3, which contains two maps. The map on the left shows the distribution of all recorded faults (those with available location information), whereas only the weather-related faults are shown on the map on the right. The size of the circles corresponds to the number of customers affected by each event and the colour to the total Customer

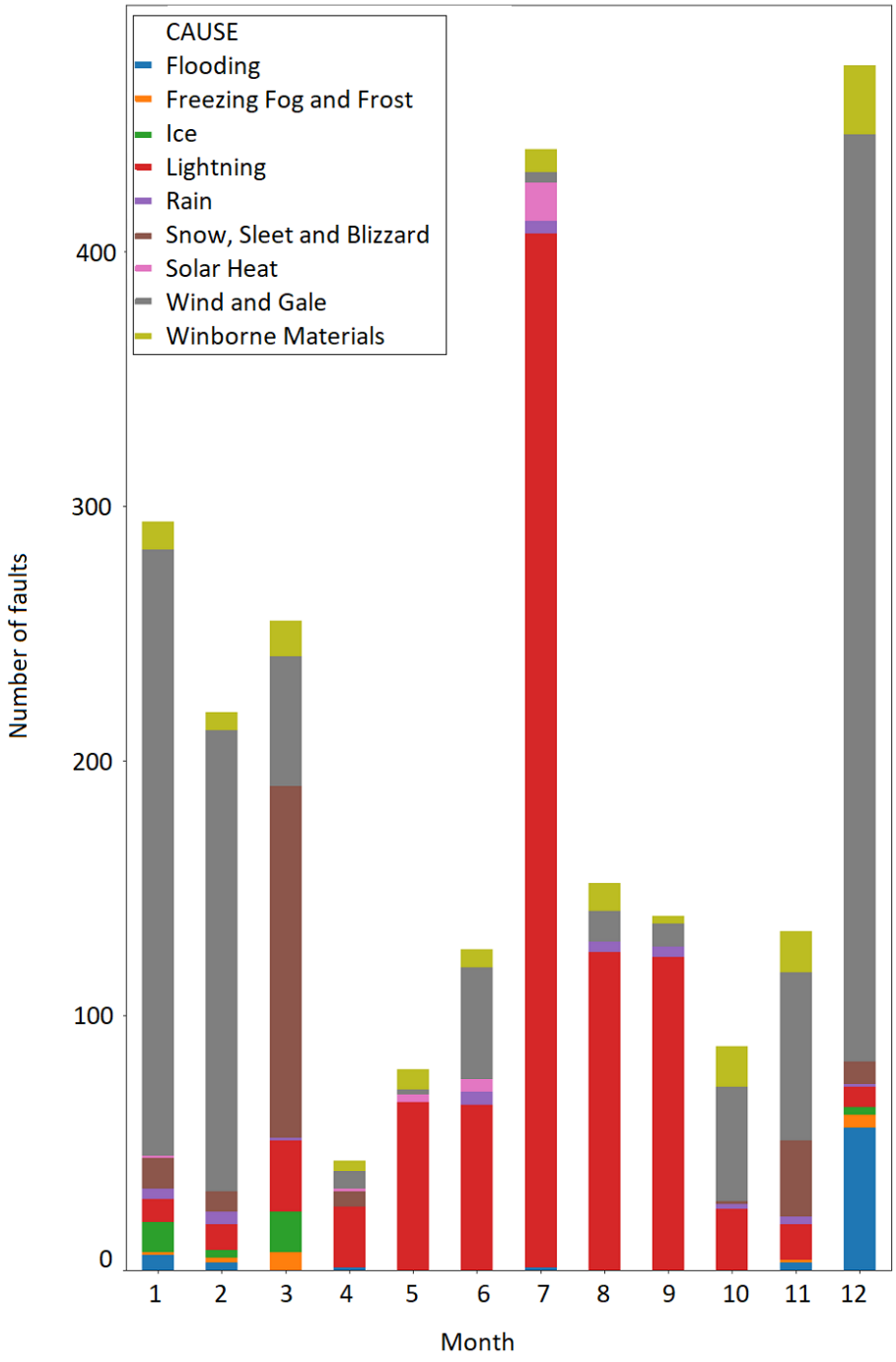


Figure 6.2: Weather related faults at the HV level by month (20/05/2013−20/07/2018).

Minutes Lost (CML). The darker colour of the circles indicates lower CML.

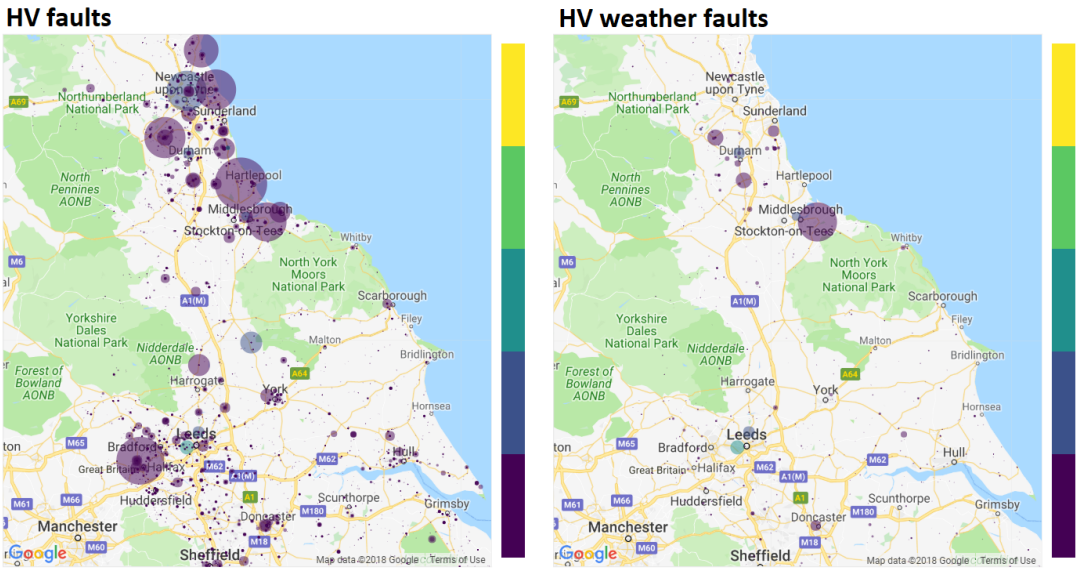


Figure 6.3: HV faults in NPG distribution network area (North East of England, Yorkshire and northern Lincolnshire) for the period 20/05/2013 – 20/07/2018. Left: all faults, Right: faults caused by weather conditions.

The LV fault records cover the period 16/06/2013 – 11/06/2018 and contain 103819 events in total. The LV faults refer to the LV side of the secondary transformer (0.4kV). The postcode of the incident location was available in the report description for 60501 of these faults, with only 711 of them being weather related faults. The number of events per weather related cause for the LV level is shown in Table 6.2. It can be seen that no faults due to "Ice" or "Freezing Fog and Frost" are present at the LV level, which is probably due to the fact that it is a predominantly underground network.

Unlike the HV level, where the numbers of weather, non-weather and unknown cause faults were comparable, this is not the case for the LV level, where the amount of faults with a registered weather cause is significantly lower than the rest. It is worth noting that even when including the 43318 LV incidents, for which no postcode was available (and were not included in the analysis), the number of events that have been registered as weather related faults is only 1034 which is still very low compared to the total amount of LV faults. This, combined with the fact that the number of unknown

Table 6.2: LV Weather Related Faults

Cause	**No. of Events**
Wind and Gale (excluding Windborne Material)	476
Rain	132
Flooding	64
Solar Heat	16
Snow, Sleet and Blizzard	10
Lightning	8
Windborne Materials	5

cause faults at the LV level is very high, could be an indication that weather related faults at LV level are underestimated as they cannot always be correctly identified, for example due to weather changes and relevant evidence having disappeared by the time a maintenance crew begins its investigation.

Similarly to the HV case, initial observations can be made from the distribution of the LV faults throughout the year and their causes, as shown in Figure 6.4.

- At the LV level the network is not affected by "Lightning" as much, as there are only a few faults that have been caused by it. However, "Rain" is the cause of a substantial number of weather related faults throughout the year.
- As is the case at the HV level, "Wind and gale" is the most prevalent cause of LV weather faults in the winter months, with December being the month with the higher number of faults due to wind and gale.

The initial observations agree with Northern Powergrid's experience that "Lightning" is one of the main causes of weather related faults at the HV level, while "Rain" has a greater impact on the LV level.

The fault occurrences and their locations for the LV level can be seen in Figure 6.5. The ratio of customers affected at HV and LV level has been taken into account so that the sizes of the circles appearing on the map are of the same order of magnitude.

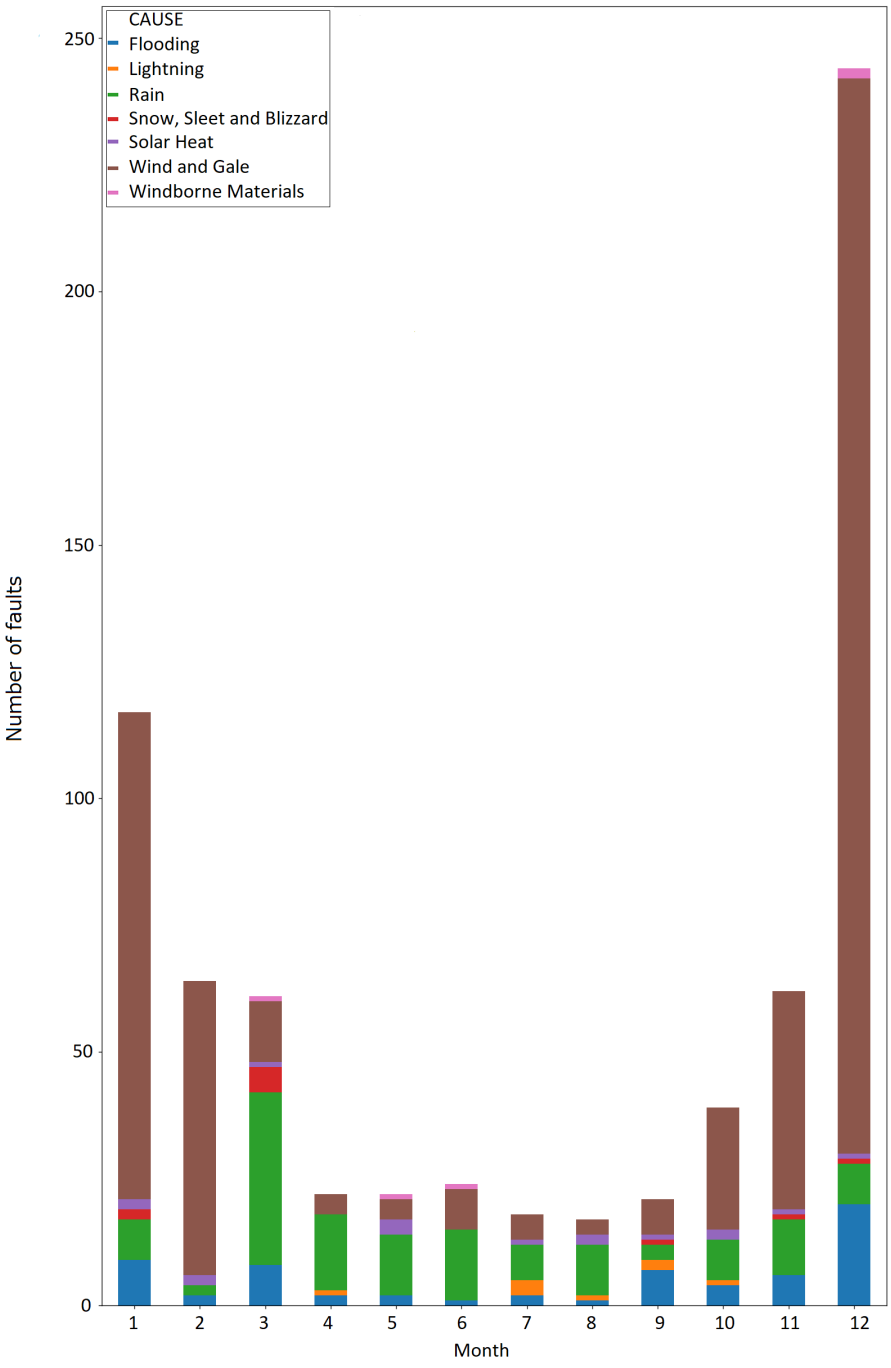


Figure 6.4: Weather related faults at the LV level by month (16/06/2013−11/06/2018).

The visualisation of the fault data on the map serves two purposes. First, it provides an easy way of assessing the impact of weather related faults and identify areas of the network that are more likely to be affected by these faults. In addition, the ratio of weather related faults with respect to the total number of faults indicates where there is a greater need for a weather related fault prediction. It is worth noting, that although the number of fault occurrences is much higher at the LV level, the HV faults are more valuable to predict as there are more customers per substation at the HV compared to the LV level. Therefore an HV fault results in higher CML.

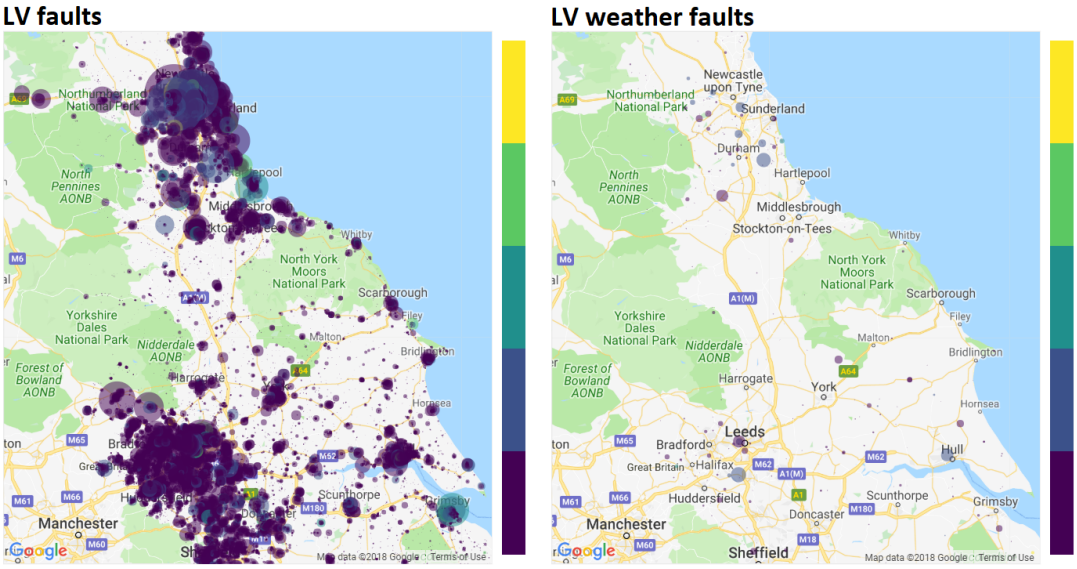
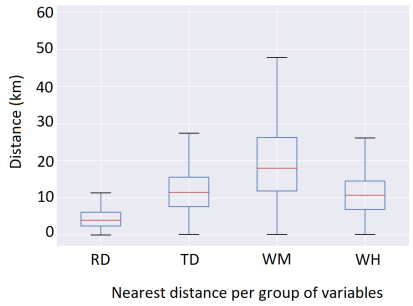


Figure 6.5: LV faults in NPG distribution network area (North East of England, Yorkshire and northern Lincolnshire) for the period 16/06/2013 – 11/06/2018. Left: all faults, Right: faults caused by weather conditions.

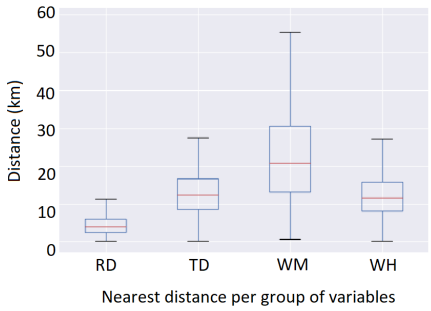
6.2.2 Weather Data

For the purposes of this analysis, access to the Met Office UK MIDAS datasets was granted by the Centre for Environmental Data Analysis (CEDA). Nineteen weather variables were considered for the analysis. The total number of active Met Office weather stations within NPG's licence area is 276. Not all measurements are available at each weather station, so data from more than one station was used to describe the weather conditions at the time of a fault. The 19 variables are categorised in 4 groups



Variable Group	Min Distance (km)	Max Distance (km)
RD	0.01	18.30
TD	0.14	34.61
WM	0.14	58.87
WH	0.14	33.34

Figure 6.6: Nearest weather station distances for the HV faults.



Variable Group	Min Distance (km)	Max Distance (km)
RD	0.19	17.90
TD	0.19	34.04
WM	0.69	58.87
WH	0.19	33.36

Figure 6.7: Nearest weather station distances for the LV faults.

of weather data: daily rainfall (RD), daily temperature (TD), hourly wind (WM) and hourly weather observations (WH). To obtain the desired data, the locations of all active weather stations within Northern Powergrid's licence area were compared to the known fault locations (postcode found in each fault's 'location text' description in the fault records) and a nearest weather station for each group of variables was assigned to each fault. For each group of weather data, the range of distances from the nearest weather station to the fault locations can be seen in the boxplots in Figures 6.6 and 6.7 for the HV and LV faults respectively. The minimum and maximum nearest distances can be found in the tables in the same figures.

To get an idea of how the weather observed to the nearest weather station relates to that of the network in question, an example showing the ambient temperature measured at the two locations is presented. Figure 6.8 shows the distance between Wooler, where monitoring data for several substations were available, and Boulmer, where the nearest

active Met Office weather station is located.

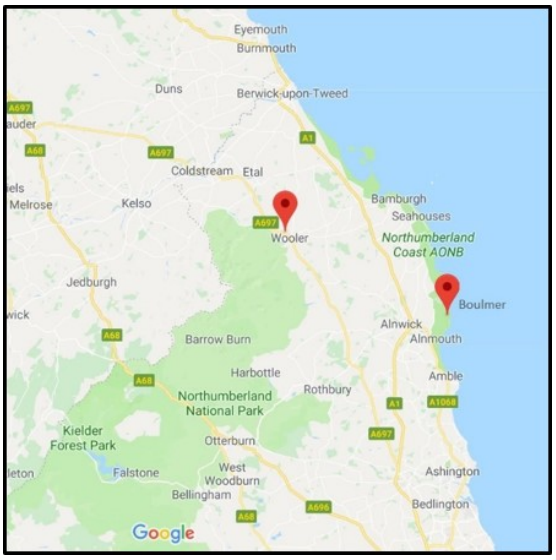


Figure 6.8: Distance between the location of a monitored substation (Wooler) and that of the nearest Met Office weather station (Boulmer).

The distance between these two locations is approximately 20 miles, which is about 32 km. It is evident from Figures 6.6 and 6.7 that this distance is almost the same as the maximum distance between the fault locations and the weather station locations for most of the weather variables considered (with the exemption of weather stations with available wind variables, where the maximum distance observed was higher). The relation between the ambient temperature of a Wooler substation, where the measurements were taken at the substation site, and the air temperature measured at the Boulmer weather station can be seen in Figure 6.9, where the line and scatter plots for these two variables for July 2015 are presented. These plots show the relation between the two sites in time (line plot) and over range (scatter plot) respectively. The two plots show that the two quantities vary in the same way, with the ambient temperature at the substation being slightly higher than the air temperature recorded at the nearest weather station. The purpose of these two plots was to identify agreement between the temperature measured at the substation and that measured at the nearest weather station. Visual inspection of the scatter plot on the right indicates that these two variables are correlated, however in order to quantify this correlation, the Pearson cor-

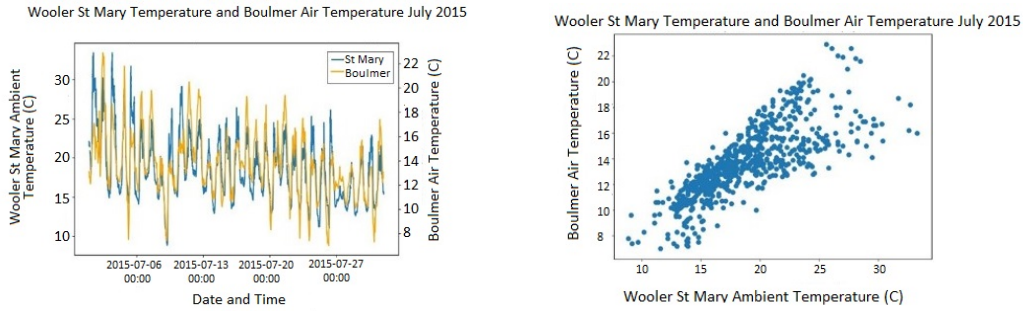


Figure 6.9: Line (left) and scatter (right) plots for the ambient temperature measured at a Wooler substation and the air temperature recorded at the nearest weather station at Boulmer in July 2015.

relation coefficient was computed as well. The value of the correlation coefficient for the temperature at the two locations was found to be 0.514, which can be considered a moderate to strong correlation. While this observation indicates a relation between the temperatures measured at the two locations it cannot be said with certainty that the general weather conditions taken from the nearest to the fault location's weather stations can accurately represent the environmental conditions at the area of the network where the fault occurred. However, there is value in exploring the use of this data for the fault prediction methodology discussed in this chapter. Due to the fact that the weather stations are not uniformly distributed across the DNO licence area, a single weather station was selected to describe the weather conditions at a given area as having two or more stations with similar distances to the fault location was very rare. Representing the weather conditions by the measurements taken at N nearest stations would require weighting the observations as a function of their distance to the fault, which would complicate the analysis whereas it is unlikely that this would increase the correlation between weather conditions and fault occurrence. The results presented in the next sections indicate that the weather observations from the nearest weather station provide a reasonable estimate of the conditions at the fault location, as the occurrence of the weather-related faults could be predicted with a good accuracy by various classifiers.

The measurements for the 19 weather variables, which are shown in Table 6.3, were taken from the nearest to each fault location relevant weather stations and were then

used to form the datasets that are discussed later in this chapter.

Table 6.3: Weather Variables (Features)

	Variable	Units	Group of Variables
1	Precipitation on the day before	mm	Daily Rainfall Measurements (RD)
2	Sum of precipitation 2 days before	mm	
3	Sum of precipitation 3 days before	mm	
4	Sum of precipitation 6 days before	mm	
5	Max daily air temperature	deg C	Daily Temperature Measurements (TD)
6	Min daily air temperature	deg C	
7	Difference in max air temperature with day before	deg C	
8	Difference in min air temperature with day before	deg C	
9	Mean wind speed	knots	Mean Wind Measurements (WM)
10	Mean wind direction	deg (true)	
11	Max gust speed	knots	
12	Max gust direction	deg (true)	
13	Air temperature	deg C	Hourly Weather Measurements (WH)
14	Dew point temperature	deg C	
15	Wet bulb temperature	deg C	
16	Humidity	%	
17	Derived hourly sunshine duration	0.1 hour	
18	Total cloud amount code	eighths	
19	Visibility	dam	

The measurements used for the variables $(1)-(8)$ were recorded on a daily basis, while hourly measurements were used for the rest of the variables. There were two reasons why daily precipitation data was selected over hourly. Based on the network owner's experience, in the case of an underground fault at both HV and LV levels, the rainfall in the days before the event has a greater impact than at the time of the event, as it takes some time for the rainfall to permeate through the ground into the cables that have been damaged by the ground movement. This is not the case for flooding events, when the faults occur fairly soon after the rain. However, since there are more Met Office weather stations collecting daily precipitation data compared to those collecting hourly, and it was possible to use data from a site nearer to the fault location, daily precipitation data was chosen for this analysis. The timescales of the selected weather variables that are used as inputs to the classification models with respect to the time of the fault (or no fault) example can be seen in Figure 6.10.

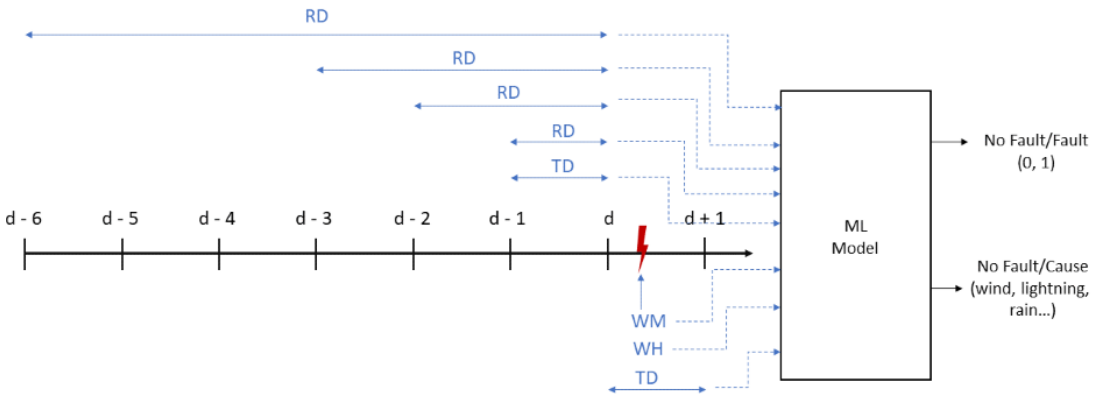


Figure 6.10: Timeline for selection of the 4 groups of weather variables approaching a fault occurrence classification. The codes RD, TD, WM and WH correspond to subsets of the weather variables shown in Table 6.3. Each group includes the following variables. RD: 1-4, TD: 5-8, WM: 9-12 and WH: 13-19.

More specifically, Figure 6.10 shows the fault that occurred some time on day d and the timescales of the variables that were chosen to describe the weather conditions associated with that fault. It can be seen that for each sample, there are some hourly measured variables that are taken at the hour of the fault and some daily variables that are taken on the days before the fault. The weather variables that correspond to each of the group of variables mentioned in Figure 6.10 can be seen in Table 6.3. Using the information available for the weather related faults and the weather measurements for the corresponding periods, a new dataset was created for each voltage level. The process of developing the new datasets is described in the next section.

6.3 New Datasets and Data Analysis Methodology

This section is divided into two subsections, which describe the dataset development and the overall data analysis process, which was presented in the methodology diagram of Figure 6.1.

6.3.1 Dataset Development

The features of the new datasets were a number of different weather variables describing the conditions for *Fault* and *No Fault* examples and were used as inputs to a number

of classifiers, which were described in the Methodology chapter. The dates and times of the *Fault* examples were taken directly from the fault records. In this analysis, anything that results in the interruption of power supply is considered a *Fault*, while a *No Fault* example can be any time when no fault occurred, i.e. there was no fault record present. As the fault records analysed in this case study include faults that occurred within a 5-year period, it would not have been feasible to include all dates and times when no fault was recorded and use these as the *No Fault* examples. Therefore, it was decided to select two *No Fault* examples for each recorded *Fault* example, with the dates and times of the *No Fault* examples chosen in a way that would ensure that the operating conditions would be similar to those of the relevant *Fault* example. The time of the first *No Fault* example was selected to be 24 hours before the fault (previous day, same time) and the time of the second *No Fault* example was one week before the fault (one week before, same day, same time). This was done to ensure that the fault was not caused by a typical time based network event[1]. If a selected date and time for a *No Fault* example coincided with a date and time of a fault in the fault records, then it was not included in the analysis. After the dates and times were finalised, the corresponding values for the selected weather variables were extracted from the nearest weather station corresponding to each of the 4 groups of variables mentioned earlier in this section.

As the measurements for the selected weather variables were not available for all dates and times in the new dataset, a number of different subsets were explored, including different combinations of weather variables and *Fault* / *No Fault* examples each time. Discarding all the data examples for which some of the weather observations were missing would lead to a small dataset, while maintaining a larger dataset size would come at the expense of features, meaning that the impact of certain weather conditions on the occurrence of faults would not have been considered. Therefore, it was decided to apply the methodology on different datasets, each having a different number of features, to see which combination of weather variables was more likely to predict the occurrence

[1] An example of a typical time-based event could be the abrupt change in load that occurs due to the daily or weekly routine of the consumers (e.g. for a network with industrial customers, the starting of heavy loads on Monday morning).

of a weather-related fault. These subsets, which had different sizes, were then used as inputs to the classifiers for a comparison in order to identify the best performing method. As explained earlier, there are a number of factors reducing the number of available data examples, including the absence of location information associated with the recorded faults and the lack of available measurements at the relevant weather station for many of those with location information. Therefore, from the initial number of 2441 recorded faults at HV and 711 at the LV level, the maximum number of faults used for prediction was 259 and 73 for the HV and LV cases respectively. The different data sets considered in this analysis are shown in Tables 6.4 and 6.9 for the HV and LV level respectively, where the dimensions (number of weather variables considered), the total number of data points as well as the *Fault* and *No Fault* examples in each dataset are presented.

The datasets resulting from the process discussed above, describe the training and test datasets that are shown in the diagram of Figure 6.1 and used in the case studies presented in the following sections.

6.3.2 Data Analysis Process

As described above, various datasets containing different combinations of weather variables were developed and used for the analysis. The Data Analysis Process part of the diagram shown in Figure 6.1 is explained below.

The main challenge is the need to accurately map environmental conditions to fault occurrence. The functional form of this relation will vary across networks, so a means of articulating it for all eventualities must possess a flexible decision surface that can be learned from past observations. To determine the optimal model choice, a selection of candidate classification techniques with diverse underlying decision surfaces were compared in order to identify the most suitable methods to classify exemplar data into: (a) *Fault* and *No Fault* and (b) *No Fault* and *Fault Type*. The classification methods applied to the data and compared are the following:

- Classification And Regression Trees (CART)
- Naive Bayes (NB)
- Logistic Regression (LR)
- Support Vector Machines (SVM)
- k-Nearest Neighbour (k-NN)
- Bagged Trees (BT)
- Random Forest (RF)
- Gradient Boost (GB)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Multi-Layer Perceptrons (MLP)
- Gaussian Process Classification (GPC)

Using a 10-fold cross validation approach, all the classifiers were applied on the same datasets and compared to find the ones that performed better. In cross validation, the dataset is split into a number of smaller subsets, which is 10 in this case. Out of these 10 subsets, 9 are used as the training set and 1 is used as the test set. This process is repeated 10 times so that all data points have been used for both training and testing. For datasets with an adequate number of faults for more than one weather related cause, the same process was used to classify the data between *No Fault* and each fault type.

As shown in Figure 6.1, this process is repeated for the different datasets in order to obtain the final results of the best performing model. Using the metrics described in Chapter 4, the overall performance of a classification method on each of the developed datasets is assessed and compared to that of the other classifiers. The 'optimal model choice' refers to selecting the model with the highest classification accuracy, precision and recall, while taking into account both the input variables (different datasets have different combination of weather variables) and the classification algorithm.

The results of the data analysis methodology described above are presented in the next section, which focuses on the best performing datasets for the cases of HV and LV faults.

6.4 Results and Discussion

The process of jointly analysing distribution network fault data and historic weather data in order to predict the occurrence of weather related faults was described in the previous sections. The results of this analysis are presented in the form of the following three case studies.

6.4.1 Weather-Related Fault Prediction at the HV level

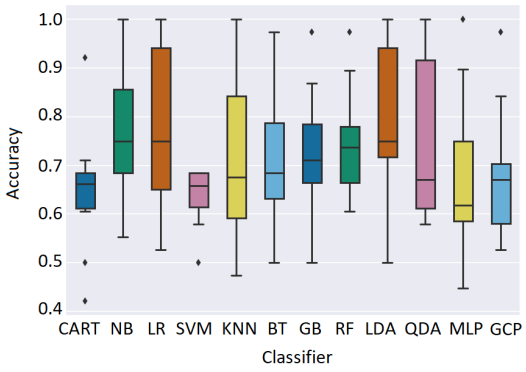
Section 6.3.1 discussed the subsets of data considered which contain a different number of *Fault / No Fault* examples and part of the weather variables shown in Table 6.3. The dataset characteristics and the accuracy of the best performing classifier for each of these subsets are summarised in Table 6.4.

Table 6.4: Summary of Results for HV Datasets

Dataset	Weather Variables	Dataset Size	Fault / No Fault	Accuracy	Classifier
#1	All	86 examples	31/55	0.728 (0.130)	RF
#2	Excluding variable (17)	277 examples	102/175	0.764 (0.174)	LDA
#3	(1) – (16)	381 examples	140/241	0.792 (0.156)	LDA
#4	(1) – (4) and (9) – (16)	717 examples	259/458	0.743 (0.165)	LDA

Note: The accuracy of prediction of the best performing classifier is represented by the mean and standard deviation of the results obtained during a 10-fold cross validation.

The numbers shown in the "Weather Variables" column in Table 6.4 correspond to the weather variable numbers in Table 6.3, the number of examples in the "Dataset Size" column is the total number of *Fault / No Fault* examples in each dataset, followed by the specific number of *Fault* and *No Fault* examples in the next column and the values in the "Accuracy" column are the mean and standard deviation (in parentheses) accuracy resulted from the cross validation process. The accuracy refers to the accuracy of prediction of a fault at a given location given the weather conditions at the time of the fault and the days before the fault. The above results show that Linear Discriminant Analysis performed better in the majority of the analysed datasets, while the highest accuracy was achieved when dataset #3, which contained 381 *Fault / No Fault* examples, was used as input to the classifier. More detailed results regarding the analysis of



Classifier	Accuracy
CART	0.648 (0.126)
NB	0.761 (0.145)
LR	0.782 (0.161)
SVM	0.638 (0.057)
KNN	0.706 (0.171)
BT	0.711 (0.134)
GB	0.719 (0.136)
RF	0.743 (0.113)
LDA	**0.792 (0.156)**
QDA	0.755 (0.162)
MLP	0.677 (0.159)
GPC	0.683 (0.130)

Figure 6.11: Cross validation results for *Fault / No Fault* classification on dataset #3 for the HV level faults.

this subset are presented below.

The 12 classifiers listed in Section 6.3.2 were compared in order to find which of them performed better in this case study. The result of this comparison for the *Fault / No Fault* classification on dataset #3 is shown in Figure 6.11, which shows the range of the fault prediction accuracies for each classifier. The box and whisker plot, which has been selected for the presentation and comparison of the classification accuracies among the different methods can be interpreted as follows. The three horizontal lines on each of the boxes correspond to the 1^{st}, 2^{nd} and 3^{rd} quartiles. This means that the bottom, middle and top lines of each box indicate that 25%, 50% and 75% of the results respectively, are below the corresponding accuracy value. The 'whiskers' extend to the 'minimum' and 'maximum' values, while the points outside this range are the 'outliers'. The 'minimum' and 'maximum' values are calculated using the interquartile range, which is the difference between the 3^{rd} and 1^{st} quartiles. The table next to the plot shows the mean and standard deviation of the resulting accuracies for each classifier for dataset #3.

The majority of faults in this dataset were caused by "Wind and Gale" or "Lightning", while only 7 faults were caused by other conditions. In order to explore the potential of the classifiers to classify the data not only into *Fault* and *No Fault* but also into the fault types, these 7 faults were removed from the dataset and the classification

process was repeated for the reduced dataset. The results of this analysis were similar to those shown in Figure 6.11, with the mean classification accuracy for the best performing classifier (LDA) being 0.792. To assess the classification results, metrics such as the precision and recall were considered alongside the classification accuracy. The expressions and meaning of each of these three metrics were given in Chapter 4.

After removing the 7 *Fault* examples mentioned above, 80% of the remaining dataset was used for training and 20% for testing. Using LDA, which was the best performing classifier, the classification results on the held-out test set (48 *No Fault* and 27 *Fault* examples) are shown in Table 6.5, which shows the confusion matrix for this classification's results. The relevant performance metrics are shown in Table 6.6.

Table 6.5: Confusion Matrix for the HV *Fault* / *No Fault* Classification

		Prediction outcome		
		No Fault	**Fault**	**Total**
Actual Class	**No Fault**	44	4	48
	Fault	6	21	27
	Total	50	25	

Table 6.6: Metrics for the HV *Fault* / *No Fault* Classification

	Precision	**Recall**	**Accuracy**
No Fault	88%	91.67%	86.67%
Fault	84%	77.78%	

It can be seen that, for this randomly chosen test set, the overall classification accuracy when using LDA is 86.67%. The model correctly classified 44 out of the 48 *No Fault* examples and 21 out of 27 *Fault* examples. Even though there are more *Fault* than *No Fault* examples that were misclassified, the accuracy is high considering the fact that only weather variables have been used to predict the occurrence of a fault. The same process was repeated in order to classify the faults based on their cause and the results

are shown in in Tables 6.7 and 6.8.

Table 6.7: Confusion Matrix for the HV *No Fault* / Fault Type Classification

		Prediction outcome			
		Lightning	**No Fault**	**Wind**	**Total**
Actual Class	**Lightning**	9	1	0	10
	No Fault	3	44	1	48
	Wind	0	5	12	17
	Total	12	50	13	

Table 6.8: Metrics for the HV *No Fault* / Fault Type Classification

	Precision	**Recall**	**Accuracy**
Lightning	75%	90%	86.67%
No Fault	88%	91.67%	
Wind	92.31%	70.59%	

Again, the overall classification accuracy is 86.67% and 44 out of the 48 *No Fault* examples have been correctly identified. Regarding the fault causes, the model correctly classified 9 out of 10 faults caused by lightning and 12 out of 17 of those caused by wind. It is worth noting that no *Fault* example was attributed to the wrong cause as all the misclassified *Faults* were classified as *No Faults*. The misclassified examples were further explored to identify if the distances of the relevant weather stations from the fault were larger than those associated with the correctly classified examples. For the data analysed in this chapter, no decorrelation distance was identified as the range of distances for the correctly classified and misclassified examples were similar. The results presented above show that it is possible to predict the occurrence of a weather related fault with a relatively high accuracy, considering only common weather variables that are not specific to a certain fault cause. From a network operator's point of view, this work could be extended to make use of weather forecasts covering their licence area in

order to identify potential fault locations ahead of time. Such an analysis could provide the opportunity for DNOs to identify vulnerable areas of their network and, therefore, be better prepared to respond to potential weather related faults.

6.4.2 Weather-Related Fault Prediction at the LV level

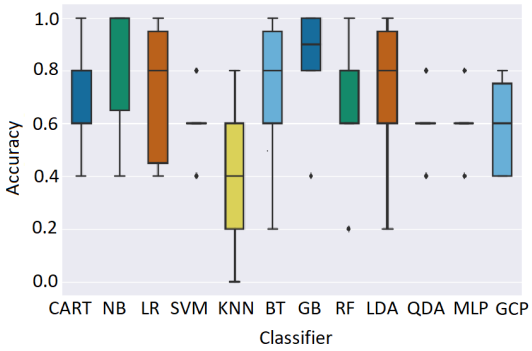
As seen in Section 6.2, the LV faults with a registered weather related cause are fewer than those at HV level, even though a much higher number of faults occurred at the LV network. This combined with the lack of location information associated with many of the LV faults resulted in considerably smaller datasets for this voltage level. The results for the best performing classifier for the LV datasets are summarised in Table 6.9. As there were only 9 *Fault / No Fault* examples with available data for all 19 weather variables, the LV equivalent to the #1 dataset (shown in Table 6.4) is not included in this table.

Table 6.9: Summary of Results for LV Datasets

Dataset	**Weather Variables**	**Dataset Size**	**Fault / No Fault**	**Accuracy**	**Classifier**
#2	Excluding variable (17)	50 examples	20/30	0.820 (0.227)	GB
#3	(1) – (16)	95 examples	36/59	0.797 (0.188)	NB
#4	(1) – (4) and (9) – (16)	205 examples	73/132	0.758 (0.183)	LR

During the LV fault analysis, the best performing dataset was found to be #2, which contained 50 *Fault / No Fault* examples. It is worth noting that the accuracy of prediction is also relatively high when datasets #3 and #4 are considered. The comparison of the results, however, indicates that the increased number of weather variables considered in dataset #2 gives a better description of the weather conditions affecting the LV network operation and, therefore, help to identify the most suitable prediction model. This is why dataset #2 was selected for a more detailed presentation of the results. When this dataset was used as input to the Gradient Boost classifier, an accuracy of 82% was achieved. The cross validation results and classifier comparison for dataset #2 are shown in Figure 6.12.

The above results show that the average classification accuracy of the Naive Bayes classifier was the same as that of Gradient Boost (82%). However, due the range of



Classifier	Accuracy
CART	0.700 (0.184)
NB	0.820 (0.244)
LR	0.720 (0.240)
SVM	0.600 (0.126)
KNN	0.400 (0.237)
BT	0.740 (0.237)
GB	**0.820** (**0.227**)
RF	0.720 (0.223)
LDA	0.720 (0.256)
QDA	0.600 (0.126)
MLP	0.600 (0.126)
GPC	0.580 (0.166)

Figure 6.12: Cross validation results for *Fault / No Fault* classification on dataset #2 for the LV level faults.

Table 6.10: Confusion Matrix for the LV *Fault / No Fault* Classification

		Prediction outcome		
		No Fault	**Fault**	**Total**
Actual Class	**No Fault**	5	1	6
	Fault	1	3	4
	Total	6	4	

accuracies during cross validation being larger in the case of Naive Bayes, Gradient Boost was selected as the best performing classifier. The dataset was randomly split into the train and test sets ($80\% - 20\%$ respectively) and the classification results on the held-out test set are shown with the confusion matrix and the performance metrics of Tables 6.10 and 6.11 respectively.

The overall classification accuracy on the test set was 80%. More specifically, 5 out of

Table 6.11: Metrics for the LV *Fault / No Fault* Classification

	Precision	**Recall**	**Accuracy**
No Fault	83.33%	83.33.67%	80%
Fault	75%	75%	

6 *No Fault* and 3 out of 4 *Fault* examples were correctly classified. Similarly to the results of the HV case study presented above, the results of this case study show that there is potential in using weather forecasts to predict the occurrence of weather related faults at the LV level as well. As discussed earlier, the number of LV faults that are attributed to weather related causes is very low compared to the total number of fault occurrences. Adopting a methodology that would successfully predict the occurrence of weather related LV faults could identify any weather related faults that would be otherwise attributed to an unknown cause. This could be another possible contribution of this analysis for the LV level as it would enable DNOs to get a better understanding of the environmental factors affecting their network. As can be seen from Table 6.2, rain and flooding are the second and third most common fault causes respectively and amount to a total of 196 out of the 711 weather related LV faults. However, for these faults, there was either no information related to their location in the fault records or no weather data available. This explains why no classification based on fault type was undertaken in this case, as almost all faults in the final datasets considered for the LV level were caused by wind.

6.4.3 Applying the Methodology on Faults of Unknown Cause

As it has been pointed out earlier in this chapter, the recorded weather related faults at the LV level were fewer than those recorded at the HV level, even though the total number of LV faults was much higher. This, combined with the fact that a very high proportion of both HV and LV faults have been attributed to an unknown cause, could be an indication that the effects of weather on faults are underestimated. To investigate this further, the methodology presented earlier in this chapter was applied to a smaller dataset, where the cause of the faults considered is unknown. The data analysis process was the same as the one described earlier. However, an additional source of fault data that was not used for the HV and LV case studies was used here. The second source of data used for the analysis, is the history of operation of a specific type of recloser devices. These devices, which replace the traditional fuses, use two fuses in parallel and when a fault occurs, they automatically switch to the secondary fuse to maintain

the power supply. The recloser operations were considered as faults and were examined alongside the recorded LV faults during the analysis period.

Unlike the HV and LV case studies presented above, where the faults occurred at any location of Northern Powergrid's distribution network and any time of the year, this case study looks into the faults that occurred at one secondary substation's distribution network in July 2015. The fault records for this substation's network showed that 6 interruptions had been recorded in July 2015. The date, start and end time for each of the recorded incidents are shown in Table 6.12.

Table 6.12: NPG Secondary Substation LV Incidents in July 2015

LV Incident Date	**Incident Start**	**Incident End**
09/07/2015	06 : 30	08 : 25
19/07/2015	13 : 00	14 : 40
21/07/2015	05 : 40	08 : 05
22/07/2015	13 : 57	14 : 20
28/07/2015	19 : 39	21 : 26
29/07/2015	04 : 50	16 : 05

Out of the 6 interruptions listed in Table 6.12, only 5 are considered for this analysis. According to the comments contained in the fault records, the incident occurred on the 22^{nd} of July was a safety interruption as engineers were on site to replace equipment. Therefore, this incident is not included in the analysis.

The recloser data for the same period contained 5 recloser operations indicating the occurrence of intermittent faults. When the recloser device data was examined, it was found that out of the 5 LV faults in July 2015, a number of recloser operations were recorded before 2 of them. The dates and times of the recloser device operations associated with the faults are given below:

1. LV incident: $28/07/2015 - 19:39:00$.

   Before that incident, 3 single recloser operations ($26/07/2015 - 06:17$, $28/07/2015 - 11:22$, $28/07/2015 - 15:34$) and 1 operation followed by a secondary rupture within 90 min (both at $18:33$) were recorded.

2. LV incident: $29/07/2015 - 04:50:00$.

   Before that incident, 2 recloser operations each followed by a secondary rupture within 90 min were recorded (all on $29/07/2015 - 03:34$).

It is evident from the recloser device data that a number of intermittent faults occurred towards the end of the month leading to the permanent faults on the 28^{th} and 29^{th} of July.

The fault activity in July 2015 (LV interruptions and recloser operations) and the precipitation amount measured at the nearest weather station is shown in Figure 6.13.

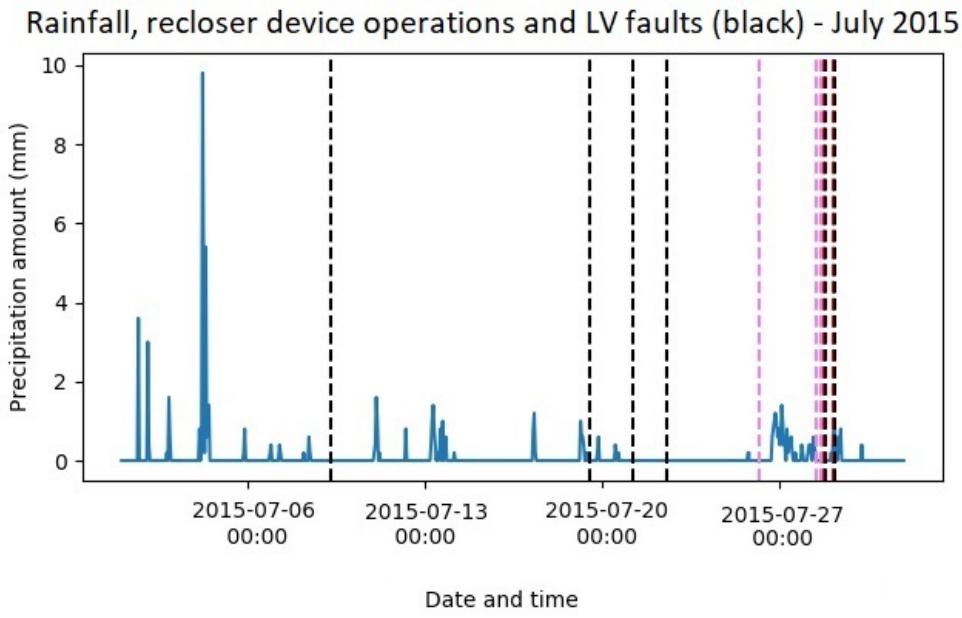


Figure 6.13: Precipitation amount (mm) for July 2015 as measured at the nearest weather station to the substation with the unknown cause faults being analysed. The faults shown in this figure are: the reclosing operations (1^{st} operation: pink, 2^{nd} operation: red) and the LV faults (black) and are plotted in the same graph.

By plotting the precipitation amount and the disturbances at the same graph, we can see that towards the end of the month there are several reclosing operations and 2 LV faults just after a period of persistent rain. Also, a few days before that, 3 LV faults occur close to each other. This could be an indication of deteriorating or faulty

equipment.

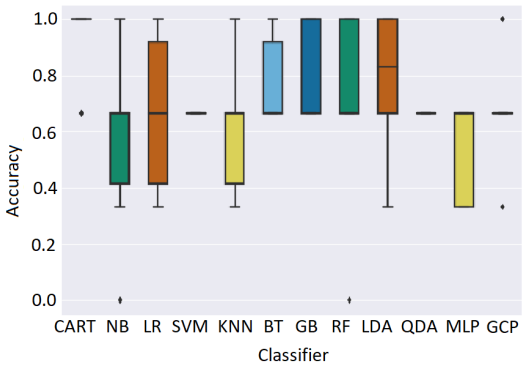
According to the fault records, the cause of all 5 LV interruptions for this period was unknown and the cause of the intermittent faults that triggered the recloser operation is also unknown. These 10 events were used as the fault examples for this case study's dataset, which consisted of 30 *Fault / No Fault* examples. As in the previous case studies, the dates and times 24 hours and 1 week before each event were used as the *No Fault* examples. Measurements for 12 out of the 19 weather variables were available for the dates and times contained in this data set. The weather variables considered correspond to the variables $(1) - (4)$ and $(9) - (16)$ from Table 6.3 and the nearest weather stations' distances from the substation are shown in Table 6.13. The dataset developed for this case study contains the same variables as datasets #4 in the previous sections.

Table 6.13: Nearest Weather Station Distances for the July 2015 Faults

Variable Group	**Distance (km)**
RD	1.38
WM	33.40
WH	6.32

The results of the cross validation process for this dataset are presented and compared in Figure 6.14, where the boxplots show the range of classification accuracy for each classifier. The mean and standard deviation accuracy of the classifiers for each dataset is given in the table.

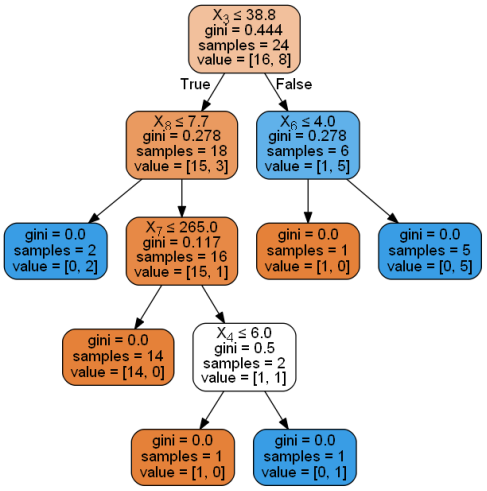
Figure 6.14 shows that the best performing classifier for the case of unknown cause faults was CART, with a mean classification accuracy of 93% resulting from the cross validation process. As in the previous case studies, this dataset was randomly split into the train and test sets $(80\% - 20\%)$ which were used to train and evaluate the CART classifier. An advantage of CART is that it can be easily visualised and interpreted. Figure 6.15 shows the decision tree learned from the train set, which was then applied



Classifier	Accuracy
CART	**0.933** (**0.133**)
NB	0.567 (0.260)
LR	0.667 (0.258)
SVM	0.667 (0.000)
KNN	0.633 (0.233)
BT	0.767 (0.153)
GB	0.867 (0.163)
RF	0.733 (0.291)
LDA	0.767 (0.260)
QDA	0.667 (0.000)
MLP	0.533 (0.163)
GPC	0.700 (0.180)

Figure 6.14: Cross validation results for *Fault / No Fault* classification on the July 2015 dataset for the unknown cause faults.

to the held-out test set and classified 4 out of 4 no fault and 1 out of 2 fault examples correctly giving an overall classification accuracy of 83%.



X_3	Precipitation 6 days before
X_8	Air temperature
X_6	Max gust speed
X_7	Max gust direction
X_4	Mean wind speed

Figure 6.15: Decision tree learned by the 80% train set from the unknown cause dataset and corresponding variables.

From the variables appearing in the decision tree, it can be seen that this classification is heavily relied on wind related variables. However, going back to the fault records, it was found that the examined part of the network is underground. While wind can have an impact on the overall weather conditions which would subsequently affect the network, it is unclear if it affects the occurrence of fault in that part of the network as much as it is implied by this decision tree. This, combined with the fact that the

Table 6.14: Comparison of Prediction Results on the Unknown Cause Faults (with and without wind variables)

Classifier	**Accuracy (including wind variables)**	**Accuracy (excluding wind variables)**
CART	0.933 (0.133)	0.700 (0.277)
NB	0.567 (0.260)	0.667 (0.149)
LR	0.667 (0.258)	0.700 (0.277)
SVM	0.667 (0.000)	0.700 (0.180)
KNN	0.633 (0.233)	0.733 (0.200)
BT	0.767 (0.153)	0.767 (0.260)
GB	**0.867** (**0.163**)	**0.867** (**0.163**)
RF	0.733 (0.291)	0.667 (0.298)
LDA	0.767 (0.260)	0.733 (0.249)
QDA	0.667 (0.000)	0.600 (0.133)
MLP	0.533 (0.163)	0.633 (0.314)
GPC	0.700 (0.180)	0.733 (0.291)

nearest weather station provided the wind measurements is much further than the weather stations with rainfall and temperature data, led to the thought of excluding the wind variables and repeat the classification process. The results for this dataset can be found alongside those of the datasets which includes the wind variables in Table 6.14.

An interesting observation can be made when comparing the results for the two datasets. The mean accuracy of CART drops significantly when the wind variables are excluded but this is not the case for Gradient Boost where the accuracy is the same for the 2 datasets (86.67%). The result of this comparison implies that the Gradient Boost result seems to be more reliable compared to that of CART. This is reasonable as CART is a simple classifier that tends to overfit the data, while Gradient Boost is an ensemble method which focuses on the difficult to predict examples. Moreover, Gradient Boost was the best performing classifier in the more general LV case study and this is a more specific LV case study, focusing on one month's fault occurrences on the same part of the network. Although this is a case study, which is dealing with a very small dataset and it is not clear how well the selected predictive model would generalise, the results presented in this case study are certainly an indication that some of these 'unknown

cause' faults could have been related to environmental factors and that the actual effect of weather on the network may be underestimated.

6.5 Conclusion

Fault prediction on networks with minimal monitoring was addressed in this chapter. After a brief discussion on the fault history of the distribution network considered and the available data, the proposed methodology towards the prediction of weather related faults using only weather data and its application on a real distribution network were presented. The results are presented with three case studies, where the performance of different classification methods on datasets with varying input variables is compared. The first two case studies involve all weather faults that occurred at the HV and LV levels. Linear Discriminant Analysis was the best performing method for weather-related fault prediction at the HV level, with an accuracy of 79.2% for both *Fault / No Fault* classification and classification based on the fault cause. For the LV level, Gradient Boost performed better in *Fault / No Fault* classification for weather-related faults with and accuracy of 82%. The above results show that it is possible to predict the occurrence of a weather-related fault at a specific part of the network using only weather variables. The final case study on a subset of unknown cause LV faults showed that there could be a higher number of faults caused by weather conditions than those registered as such, as prediction results on the unknown cause dataset were also good, although fewer weather variables were considered for that case. This could also be used to retrospectively analyse persistent faults with unknown causes and potentially attribute them to certain weather conditions.

The contribution and novelty of this work is a methodology for finding the functional relation between fault occurrence and environmental conditions. A practical use case stemming from this methodology would be using the model with a longer term weather forecast to understand which parts of the network were at risk of fault under the forecasted weather conditions. This would assist in the refinement of spares budgets and strategic positioning of maintenance staff although at shorter timescales. With

this use case in mind, the benefits for network operators could be further enhanced by moving the methodology towards a probabilistic framework which would in turn accommodate uncertainties in forecasts and measurement errors to provide probability of fault rather than just prediction. As distribution network operators face increasingly diverse challenges on their ageing infrastructure, such an approach would allow them to act on predictions according to their attitude to risk which in turn could be informed by asset health and criticality indices.

Chapter 7

Substation Duty Cycle Impact on Distribution Fault Occurrence

Discussions with DNOs have revealed that a high number of faults occur in their distribution network when there is a transition from spring to summer or summer to autumn. This observation could be an indication that the changing substation duty cycles affect the network in ways that have not yet been understood. As discussed in Chapter 2, load profile clustering is an area that has been thoroughly researched in order to identify repeating patterns in substation load behaviour. However, little has been done to assess the impact that the recurring load profiles and the transitions between them have on the distribution network. The work presented in this chapter investigates the relationship between the changing load profiles and the occurrence of faults or power quality disturbances. The data used, the methodology that was followed and the results are presented and discussed in the following sections.

7.1 Data and Methodology

For the research work presented here, monitoring data from the distribution network of Electricity North West, which is one of the UK's DNOs, was utilised. The data was collected from ENW's distribution network during the LCNF project Capacity to

Customers, which run from January 2012 to December 2014. PQubes, which are the power quality monitoring devices introduced in Chapter 5, were used to gather the data. For this piece of work, only the 5-min average current across the 3 phases measured at the LV side of secondary substations and the recorded power quality events were utilised.

The work presented in this chapter has 4 stages: Data Processing, Clustering, Visualisation and Event Day Prediction.

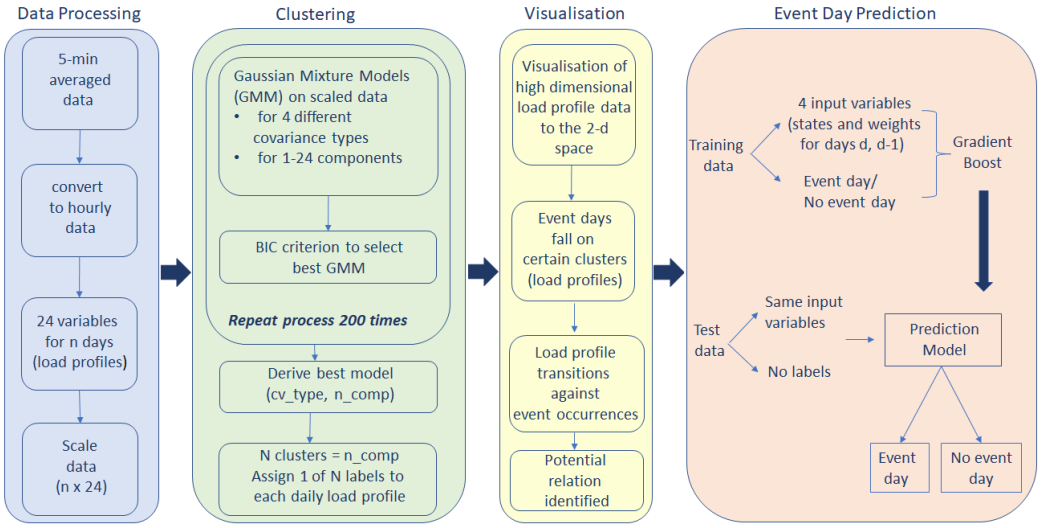


Figure 7.1: Data analysis methodology for each substation with n days of data.

Unlike the weather fault prediction case study of Chapter 6, which focused on the fault prediction stage, this chapter's case study involves all stages of the proposed Methodology, which was presented in Chapter 4. After the required preprocessing of data, this chapter's analysis begins with the 'Clustering' stage which corresponds to the 'Network Behaviour Characterisation' stage of Chapter 4 and then moves to the 'Visualisation' stage, which corresponds to 'Anomaly Detection' as a potential relation between load behaviour and event occurrence is identified in this stage. Finally, the 'Event Day Prediction' corresponds to the 'Prediction of Faults and Disturbances'. To show how this chapter's methodology relates to that of Chapter 4, as described in

Figure 4.1, the same colours have been used in the two diagrams. The steps of the analysis involved in each of the stages can be summarised in Figure 7.1. These 4 stages are discussed in detail in the following subsections.

7.1.1 Data Processing

The first stage of the analysis involved all the necessary steps in order to bring the data in the final form that would be used as input to the clustering algorithm of the second stage. For each of the substations considered, the 5-min data was initially converted to hourly data. These hourly values were then used as the current values for 24 variables, representing the daily load current profiles for n days, where n is the number of days, for which monitoring data was available and is different for each substation. The last step of data processing was to scale the data so that the values for each feature would look like normally distributed data. The size of the resulting dataset is n x 24, meaning that it has n data points (days) and 24 dimensions (variables).

7.1.2 Clustering of Load Profile Data

The scaled data obtained from the first stage, was used as input to the clustering stage. The clustering algorithm that was used to identify repeating load profiles within the data was Gaussian Mixture Models, which was described in the Methodology (Chapter 4). The reasons for selecting GMMs were discussed in Chapter 4. Among other reasons it was mentioned that the load profile data follow a multi-modal distribution, which makes GMM a good model choice. For one of the substations considered (substation #2), the probability density function for each of the 24 features and their multi-modal shape is shown in Figure 7.2.

A number of different values were tested for the 'covariance type' and 'number of components' parameters and the Bayesian Information Criterion (BIC) [95] was used to select the best model. The covariance types that were considered in this analysis and their meaning can be found in Table 7.1.

The number of components of a GMM can be known in advance or inferred from the

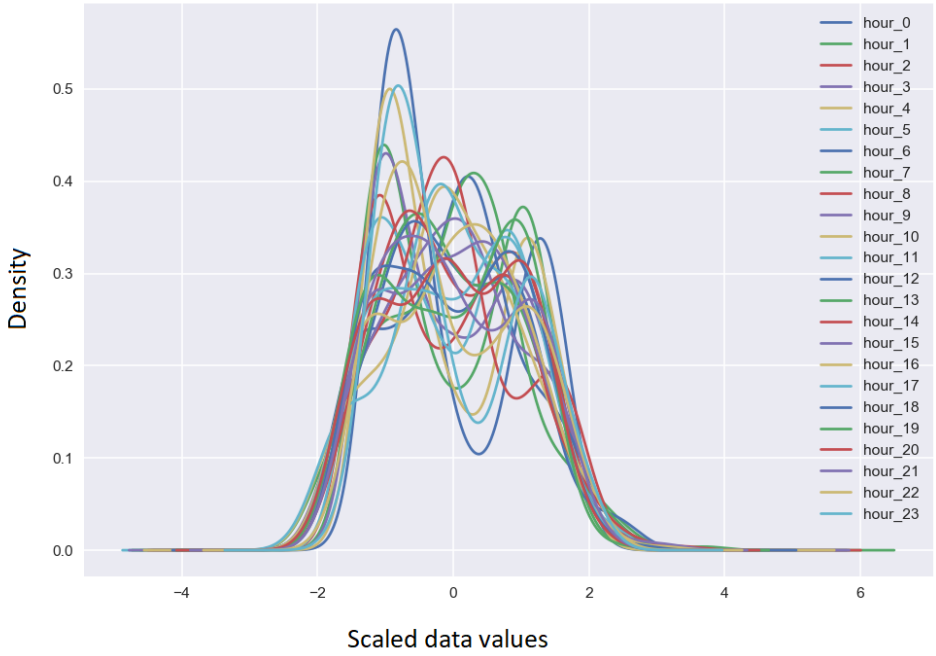


Figure 7.2: Probability density function for the 24 features for substation #2.

data based on the application. Here, the number of components, which represents the recurrent load profiles, is selected after a comparison of different models. A number of models with different parameter values were compared and the Bayesian Information Criterion was used to select the best model. This criterion is used to select the best performing model among a finite number of models. The preferred model is that with the lower value for BIC, which is defined as:

$$BIC = -2LL + k\dot{\log}(n) \tag{7.1}$$

where LL is the model's log likelihood, k the number of parameters and n the sample size.

After 200 replications of the above process, with different model parameterisations, the best model, namely the optimal covariance type and number of components, was derived. The best model was then applied on the substation load profiles, which were

Table 7.1: Covariance Types and Their Meanings

Covariance Type	**Meaning**
Full	Each component has its own general covariance matrix
Tied	All components share the same general covariance matrix
Diagonal	Each component has its own diagonal covariance matrix
Spherical	Each dimension has the same variance

subsequently grouped into N clusters of different load behaviours. The final step of this stage involved the assignment of 1 of N labels (cluster numbers) to each daily load profile.

7.1.3 Data Visualisation

To visualise the high-dimensional data considered in this analysis, three different methods were used and then compared. Each high-dimensional data point in this case is a load profile with each hour of day being represented by one dimension. First, the visualisation of the daily load profile data and the mixtures derived from the application of GMM was a simple scatter plot of the first two variables. In our case, these variables are the average currents for the first and second hours of each day. The purpose of this visualisation is to get an initial idea of any clusters formed in the data by looking into how the load changes between these two hours for each day for the duration of the available data. However, as this method takes into account only 2 out of the 24 variables significant information regarding the structure of the data remains hidden. To get a better understanding of the load profile data, identify any recurrent patterns and select the best way to visualise this data, two dimensionality reduction techniques, which take into account all 24 variables, are also used and compared in this chapter. PCA and t-SNE, which can take the initial high-dimensional data and transform it to the 2-dimensional space, are the other two methods that were used to visualise the data. In PCA, the principal components of the data are uncorrelated variables which are selected in a way that as much of the information contained in the initial correlated variables is preserved, while in t-SNE a mapping of the high-dimensional data in a 2-dimensional space which preserves the local structure of the initial data is performed.

More detailed description of these methods was given in Chapter 4.

7.1.4 State Transition - Event Day Prediction

After the best model was identified and applied to the substations' data, the N representative load profiles for each substation were generated and used to label the substation days. Then, the visualisation methods discussed above were used to view and assess the results. Subsequently, the potential effect of state transition (the transition from a daily load profile to another) on the occurrence of certain power quality events was investigated. Using the same plots to examine each substation's state transitions and the occurrence of power quality events, the load behaviour prior to an event occurrence was explored and compared to the behaviour corresponding to the rest of the time. It was observed that, for a number of substations, certain event types occurred when there were specific transitions between the load profiles. Based on this observation, the prediction of days with certain events using a predictive model that is used to identify the relation between certain load behaviour and the occurrence of power quality disturbances was investigated. The machine learning algorithm used in this stage was Gradient Boost, which has been described and discussed in the previous chapters. The input variables used to predict an event day include the load profile on the specific day (referred to as state on day d), the load profile on the day before (state on day $d-1$) as well as the weights of the mixtures that correspond to the two states, as derived from the application of the GMM. Therefore, the datasets fed into the predictive model had 4 dimensions and n data points, where n is the number of available days for each substation as shown in Tables 7.2, 7.3, 7.4.

7.2 Collective Results

The complete dataset that was used for this analysis, contained data for 75 secondary substations across ENW's distribution network. Out of the 75 substations, 16 were omitted as there were either very few days with monitoring data or very few days with power quality event occurrences. Data for the remaining 59 substations were

used for the analysis presented in this chapter. Initially, load current data from 59 substations were collectively analysed and all power quality events were considered. The methodology presented in the previous section was applied to each substation and the results were assessed using the prediction accuracy and the precision and recall metrics for both event and no event days.

7.2.1 All Power Quality Events Considered – 59 Substations

First, all 59 substations and all power quality events recorded were considered. This case takes all event types in order to get an initial idea of the range of prevalence, where relatively good prediction results were achieved. The term prevalence refers to the ratio of event days over total substation days. As this part considers all different event types, which can be triggered by different operating conditions, it is reasonable that the prediction results may not be as accurate as in the case where certain event types are considered. However, as the purpose of this part was to identify an initial range of prevalence where the prediction model performed well, the threshold above which the results were considered satisfactory were not set very high. As what is considered a 'good' accuracy can be subjective and application specific, the primary reason for making this distinction here was to identify those substations, where the proposed methodology performed better and the occurrence of power quality events was more likely to be predicted. For this case, 'good' results were those where the prediction accuracy and at least 3 out of the 4 other metrics were above 70%, while the accuracy for the 'relative good' results was above 60% and no metric was below 40%. Using these thresholds, 11 substations showed good or relatively good results, which are shown in Table 7.2.

It can be seen that the prevalence of events on these substations ranges between $0.18 - 0.77$. With this range of prevalence values in mind, the same process was repeated for specific event types. The most commonly recorded power quality events in the dataset being analysed were found to be phase overcurrent and voltage swell. The observed high prevalence of these two types of power quality events indicate that these events could be seen as a nuisance and, therefore, it was decided to focus on these two event

Table 7.2: Substations With Selected Results (All Events)

Substation number	Accuracy	Precision (Event)	Recall (Event)	Precision (No Event)	Recall (No Event)	No. of Days (Samples)	Prevalence	Results
1	0.896	0.855	0.935	0.913	0.883	229	0.6	good
2	0.821	0.878	0.74	0.833	0.8	327	0.36	good
3	0.738	0.723	0.757	0.7	0.77	232	0.54	good
4	0.778	0.708	0.811	0.613	0.865	453	0.65	good
5	0.715	0.762	0.695	0.574	0.838	435	0.53	good
8	0.885	0.965	0.552	0.904	0.76	80	0.18	good
14	0.806	0.634	0.844	0.437	0.918	347	0.77	relatively good
22	0.767	0.83	0.567	0.865	0.485	448	0.25	relatively good
26	0.617	0.552	0.722	0.743	0.516	462	0.56	relatively good
28	0.695	0.756	0.517	0.828	0.401	456	0.30	relatively good
30	0.671	0.639	0.703	0.661	0.679	458	0.54	relatively good

types. In addition, both of these event types can be the result of a power system fault. Among other reasons, an overcurrent event might occur as a result of an incipient fault which can, in time, develop into a permanent fault, while a single line to ground fault can lead to a voltage swell on the unfaulted phases. The prediction results for all substations with adequate number of phase overcurrent (34 substations) or voltage swell (27 substations) events are given and discussed below. No substations with fewer than 20 event days (for each event type) were considered for this analysis.

7.2.2 Phase Overcurrent Events Only – 34 Substations

Out of the 59 substations initially considered, 34 were used for this analysis, as these had recorded phase overcurrent events on 20 or more days. For this case, only three substations showed good prediction results, which are shown in Table 7.3.

Table 7.3: Substations With Selected Results (Phase Overcurrent Events)

Substation number	Accuracy	Precision (Event)	Recall (Event)	Precision (No Event)	Recall (No Event)	No. of Days (Samples)	Prevalence	Results
1	0.889	0.846	0.93	0.906	0.876	229	0.6	good
3	0.781	0.858	0.698	0.773	0.793	232	0.40	good
4	0.787	0.769	0.819	0.863	0.7	453	0.46	good

Looking at Table 7.3, it can be seen that the prevalence values of event days for the three substations were 0.6, 0.4 and 0.46, which are within the range $0.18-0.77$ identified earlier in this section. Out of the remaining substations, which did not give good results, only 5 had prevalence values in that range. It is worth noting that the prevalence for

all 5 of them was between $0.19 - 0.25$. The prevalence for the rest of the substations was outside this range as there were very few phase overcurrent events recorded in all of these substations, compared to the available days of data.

7.2.3 Voltage Swell Events Only – 27 Substations

Similarly to the previous case, 27 substations were selected for this part of the analysis as they had 20 or more days with voltage swell events. Out of these, 11 substations showed good prediction results. The results are given in Table 7.4.

Table 7.4: Substations With Selected Results (Voltage Swell Events)

Substation number	Accuracy	Precision (Event)	Recall (Event)	Precision (No Event)	Recall (No Event)	No. of Days (Samples)	Prevalence	Results
2	0.822	0.889	0.728	0.831	0.806	327	0.35	good
4	0.683	0.688	0.684	0.532	0.804	453	0.55	relatively good
5	0.698	0.762	0.656	0.623	0.781	435	0.48	relatively good
14	0.841	0.855	0.839	0.703	0.925	347	0.63	good
18	0.693	0.644	0.747	0.711	0.678	439	0.55	relatively good
21	0.739	0.725	0.757	0.752	0.726	424	0.51	good
23	0.705	0.76	0.613	0.771	0.594	469	0.37	relatively good
26	0.925	0.928	0.917	0.978	0.756	462	0.25	good
29	0.707	0.744	0.649	0.781	0.596	477	0.39	relatively good
30	0.728	0.793	0.659	0.723	0.734	458	0.43	good
43	0.966	0.973	0.896	0.99	0.739	476	0.11	good

The prevalence for 10 out of the 11 substations shown in Table 7.4 have a value in the range $0.18 - 0.77$, while there is one substation with good results and prevalence 0.11. In addition, there were 4 substations within the above range which did not give good results and the prevalence values for these substations were 0.19, 0.22, 0.26 and 0.72. All of these values are very close to the lower or upper limits of the selected range. The remaining 16 substations had prevalence values outside this range and, with the exception of substation #43, did not give good results when the methodology was applied to them.

When looking at the presentation of how the methodology was applied to a number of substations and the collective results discussed above the following observations can be made. First, it is evident that when there were enough days with genuine power quality events to properly train the models, the days with events could be predicted with a high accuracy. In addition, better prediction performance was observed when

considering single event types compared to the case were all event types were taken into account. This is reasonable as different event types can be triggered by different changes in the load behaviour. Lastly, for the substations with 'bad' performance, for which the value of prevalence was within the selected range, that value was found to be very close to the upper or lower limits of the range and this was true for all relevant substation in both cases, where single type events were considered.

7.3 Prediction of Days with Voltage Swell Events

In this subsection, a detailed presentation of the data analysis results, covering all steps of the methodology as described in Section 7.1, is given. The analysis results presented here were obtained from the application of the methodology to substation #2 for the case of voltage swell events.

The duration of available measurements for substation #2 was 327 days, out of which 113 days had one or more voltage swell events recorded, meaning that the prevalence of event days for this substation was $113/327 = 0.35$. After the initial processing of data, the daily load profiles corresponding to each of the 327 days were obtained. These load profiles were used as inputs to a Gaussian Mixture Model and BIC was used to identify the optimal parameters (covariance type and number of components) as discussed in Section 7.1. This process was repeated 200 times in order to obtain the best GMM model for this substation, which was found to be a GMM with tied covariance and 10 components (mixtures), which are shown in Figure 7.3.

Figure 7.3 shows the 10 representative load profiles for substation #2 as they were produced by the GMM with the above parameters. The x-axis of the diagrams corresponds to the hour of day, while the y-axis is the mean current drawn, in amperes (A). After the representative profiles were identified, the next step was to label all 327 daily load profiles, i.e. to assign one representative profile to each of the days with available data for this substation. The three visualisation methods discussed earlier in this paper were subsequently used to visualise the labelled data. The visualisation results are shown in Figure 7.4, where the points in each of the three scatter plots correspond to the 327

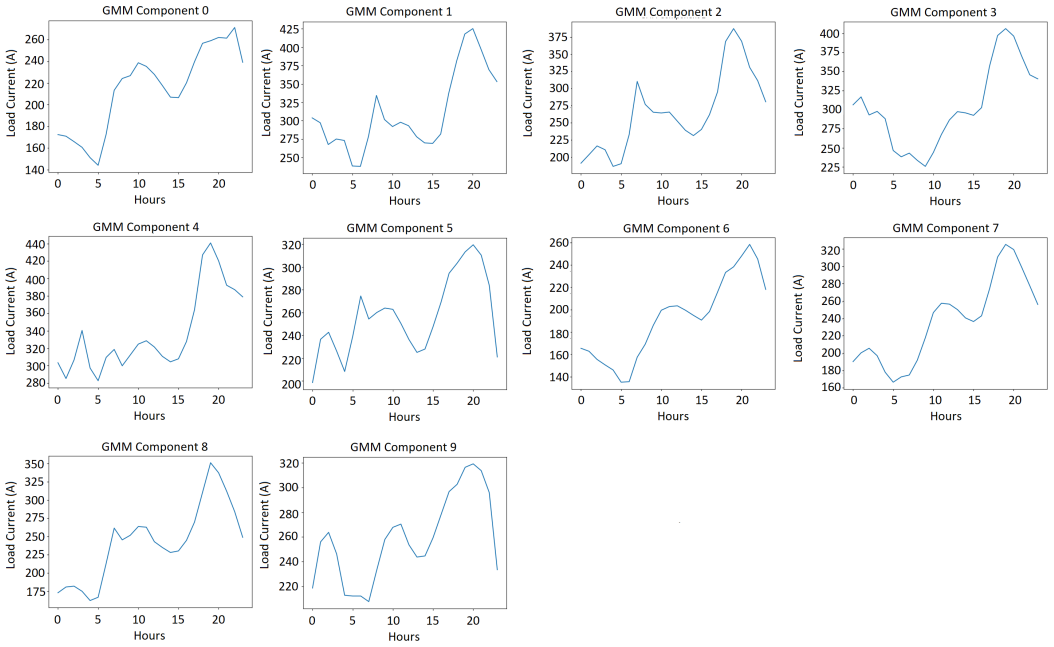


Figure 7.3: Load profiles obtained as GMM means for substation #2.

substation days.

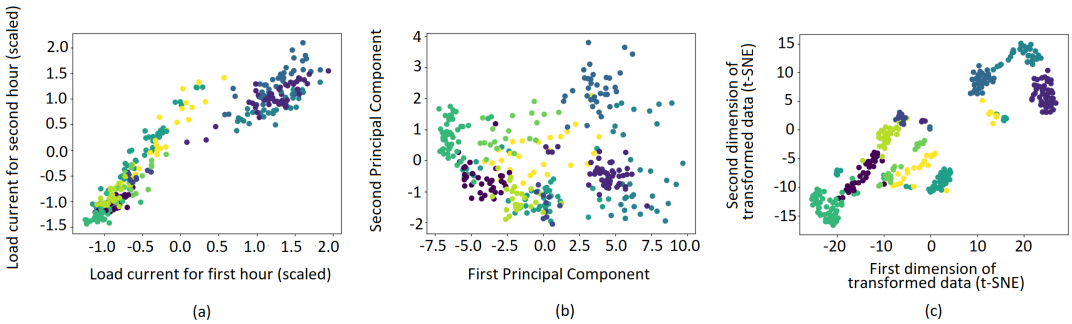


Figure 7.4: Visualisation of results : (a) scatter plot of the first two variables (scaled load current for the first and second hour), (b) PCA and (c) t-SNE.

Figure 7.4(a) shows the 2 first (out of the 24) variables, meaning that this is a plot of the first against the second hour of the day for each of the 327 days. For the plots of Figures 7.4(b) and 7.4(c), the dimensionality reduction methods PCA and t-SNE respectively have been applied to all 24 variables in order to find a projection of the 24 point load profiles into the 2-dimensional space. It is clear that t-SNE is the preferred

method of visualisation for this application as the natural spatial separation facilitated by t-SNE corresponds to the partitioning specified by the GMM clustering. This is evident from Figure 7.4 as t-SNE has achieved almost perfect grouping of the days with the same load profiles, as opposed to the other two visualisation methods. The different clusters can be seen more clearly in Figure 7.5 , which is an annotated version of 7.4(c).

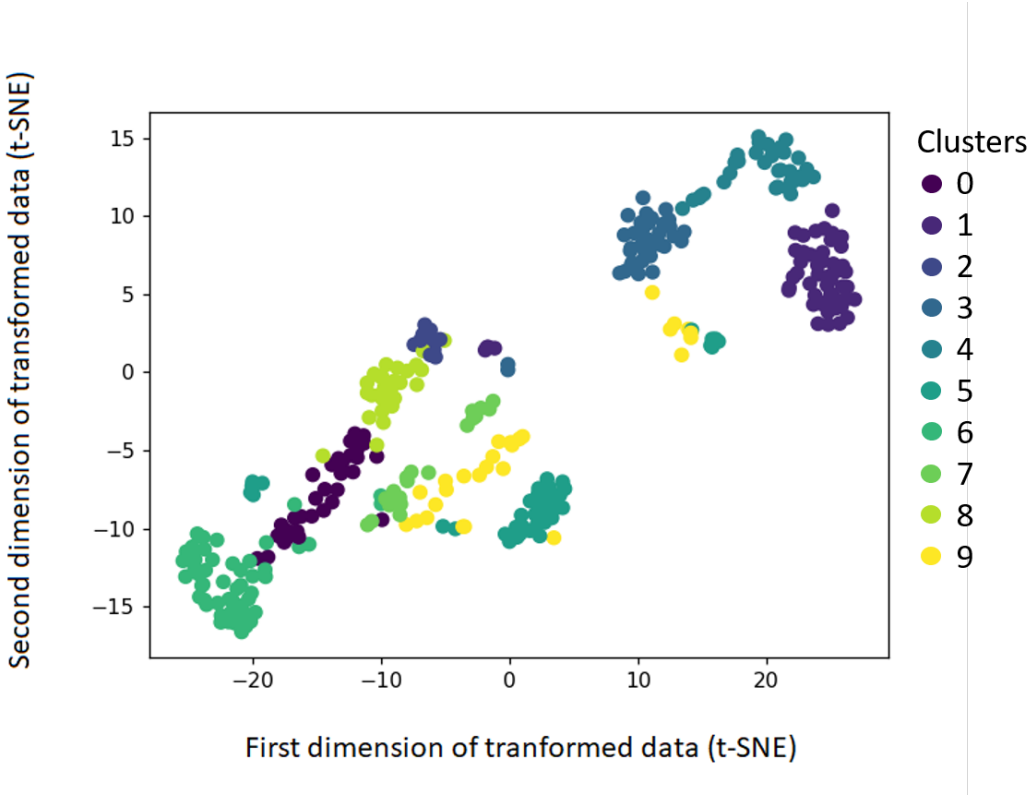


Figure 7.5: t-SNE visualisation of the partitioning achieved by GMM for substation #2, coloured by the cluster label.

When the scatter plot of Figure 7.5 was coloured based on the occurrence of a voltage swell event on a day, it was found that the event days were concentrated in specific parts of the plot. This can be seen in Figure 7.6, where the red colour indicates the occurrence of a voltage swell event, while the blue data points correspond to days with no voltage swell event recorded.

Indeed, the vast majority of recorded events seem to occur when there is a transition

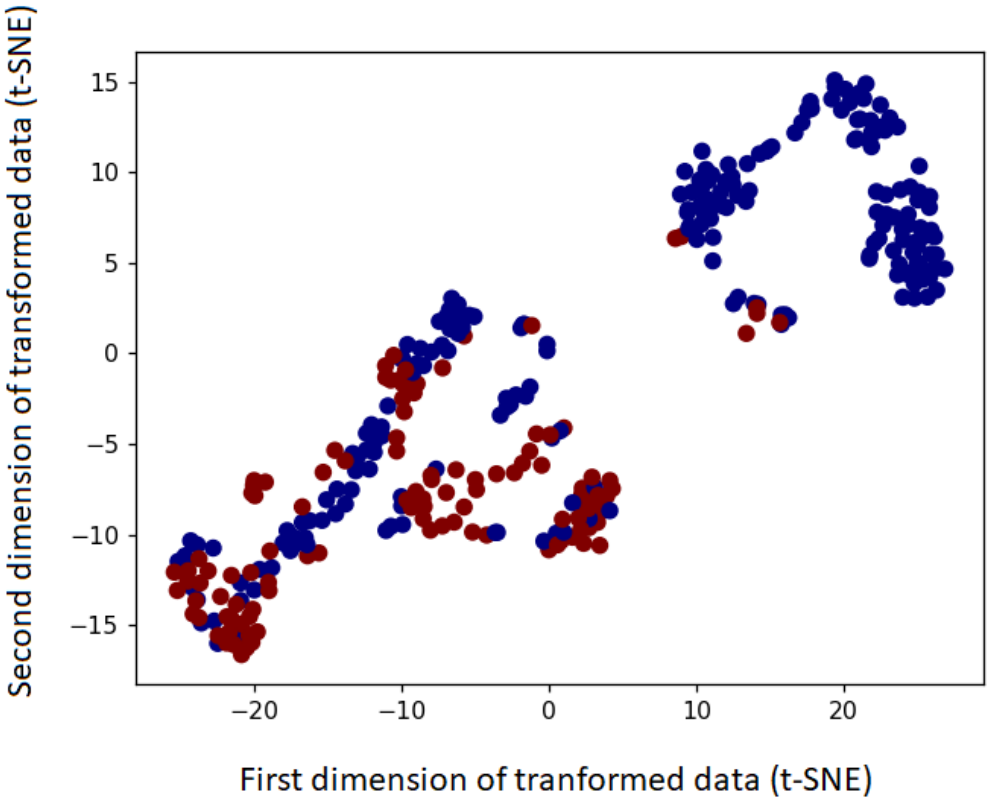


Figure 7.6: t-sne visualisation for substation #2 days – red: voltage swell event days, blue: days with no voltage swell events.

between certain load profiles or when the load dwells in a specific state. This can be seen in Figure 7.7, where the orange line shows the load profile changes and the blue line shows the number of voltage swell events on a specific day.

In order to assess the effect of changing load behaviour on the occurrence of voltage swell events, Gradient Boost was used to see how well it could predict the occurrence of a voltage swell event on a specific day, if the load profile on that day (state d) and the day before (state $d-1$), as well as the weights of the two states are known. To do this, 80% of the 327 substation days were used for training and 20% for testing. The process was repeated 100 times, where the training and test data resulted from random splits every time. These splits resulted in different days being used for training of the model and as it would be expected the results of each of the 100 replications were not

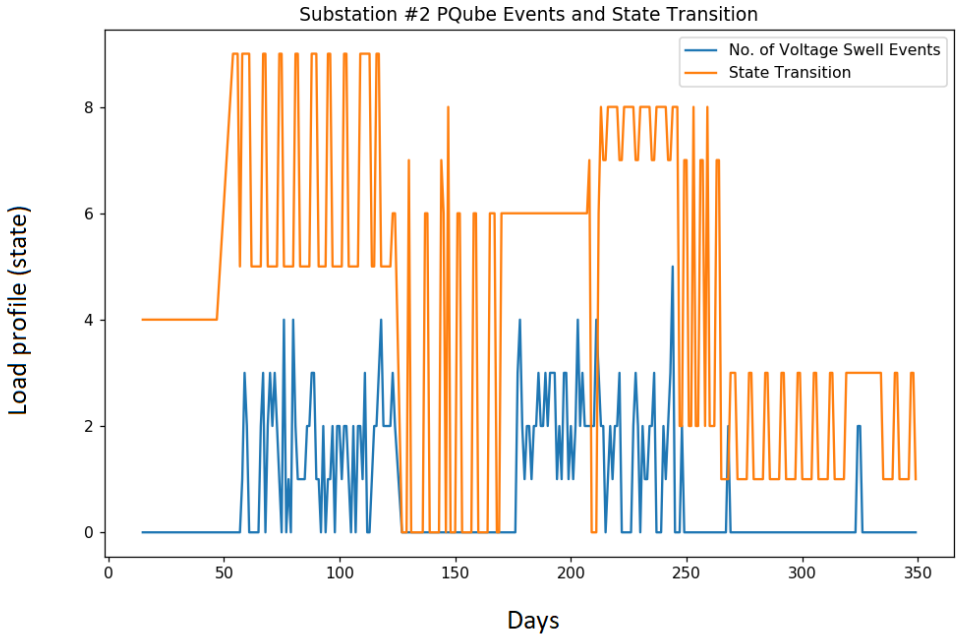


Figure 7.7: State transitions and voltage swell occurrences.

identical. The results are given in Table 7.5 as the mean value over 100 replications for each of the performance metrics.

Table 7.5: Gradient Boost Prediction Results For Substation #2

Mean Accuracy of 100 replications	0.833
Mean Precision (Event) of 100 replications	0.717
Mean Recall (Event) of 100 replications	0.890
Mean Precision (No Event) of 100 replications	0.932
Mean Recall (No Event) of 100 replications	0.801

As expected, the above results are not identical with those presented in Table 7.4 for substation #2, which refer to a single random split of the data for the same substation. However, given the fact that the results of Table 7.5 show the mean performance of 100 replications, these results indicate that days with certain power quality events can be predicted if the expected load behaviour on that day and that of the previous day are known.

7.4 Applying Methodology to Multiple Substations

The previous sections discussed the results obtained from applying this chapter's methodology on individual substations and it was found that it is possible to predict the occurrence of power quality events at certain substations using these substations' representative load profiles, when these profiles were obtained from the application of a GMM per substation. In this section, the application of a GMM to load current data of multiple substations, in order to obtain the representative load profiles across many substations, and the subsequent prediction of power quality events are explored.

After the methodology was applied to all 75 substations (including those with few or no event occurrences), the representative load profiles across these substations were obtained. The total number of substation days of the final dataset containing data from all 75 different substations was 24120, which implied that a higher number of representative load profiles would be obtained compared to those of a single substation. Initially, this chapter's methodology was applied to the whole data set. However, as the number of substation days is very high and the comparison of various GMMs in order to determine the required parameters is a very time consuming task, a subset of 2500 randomly selected days (which are just above 10% of the data) belonging to all substations was used to identified the best performing model. As discussed earlier in this chapter, BIC was used to select the optimal number of components (load profiles) from a range of components between 20 and 90. The selection process was repeated 22 times and the BIC curve for each repetition is shown in Figure 7.8. The blue dots in the graphs correspond to the BIC values of each of the models (with number of components from 20 to 90) and the orange curve is the best fit quadratic curve for these points. It can be seen that for all repetitions the BIC value decreases as the number of components increases up to approximately 60 and then it starts to increase again.

When all of the above BIC values are plotted in the same figure, the relevant best fit curve is shown in Figure 7.9. The start of the curve (which corresponds to the model with 20 components) is marked with the first red point, while the second red point

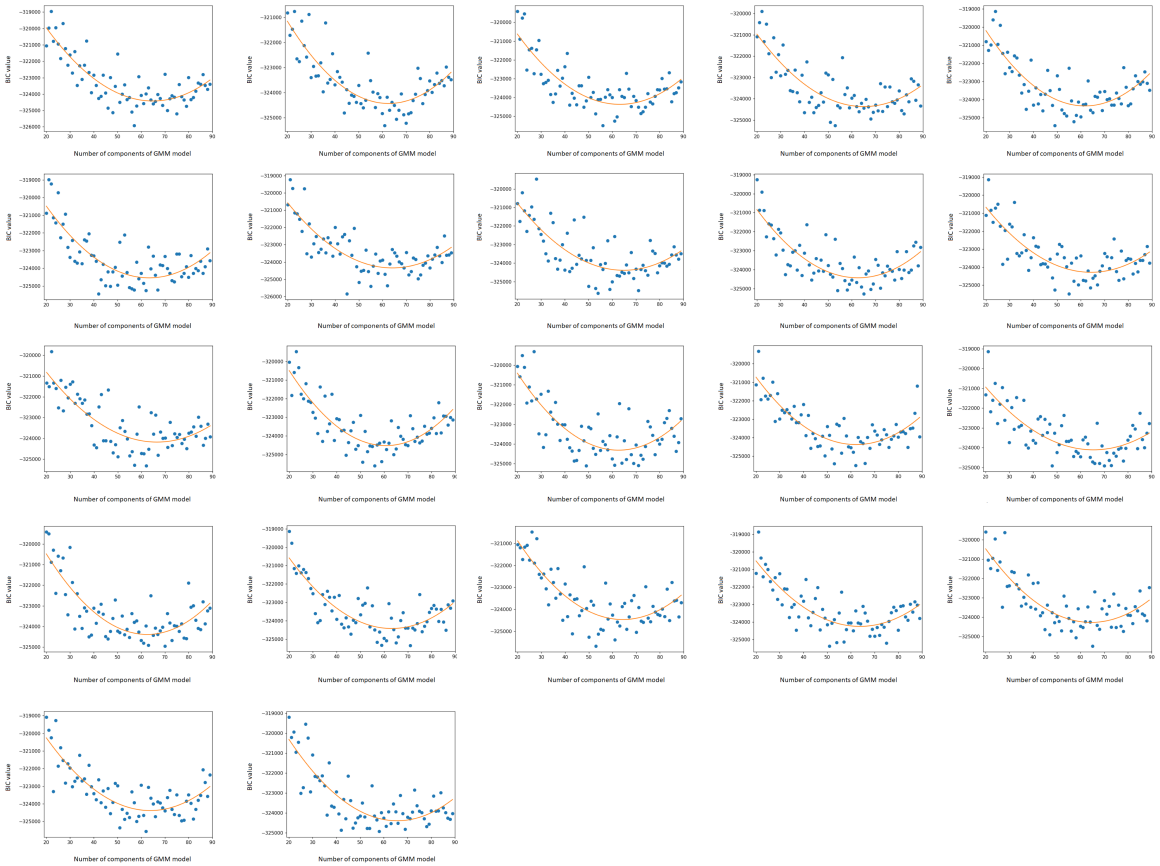


Figure 7.8: BIC curves to select the optimal number of components for each of the 22 repetitions.For each subfigure, the x and y axes show the number of components and the y value respectively.

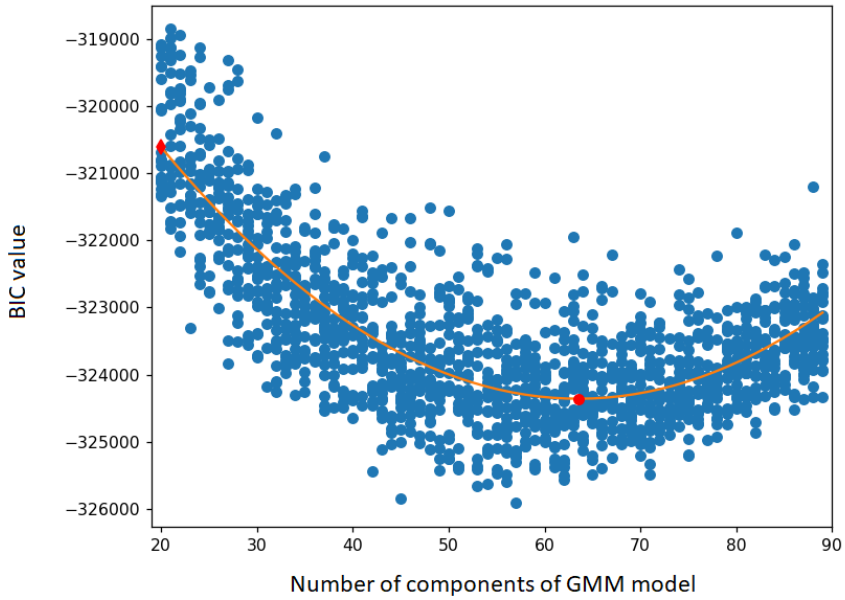


Figure 7.9: Best fit curve for all BIC values to select the optimal number of components for the GMM.

indicates the number of components which minimises the curve. Therefore, the GMM which better clusters the load profiles of the group of substations was found to have 63 components. This curve, and the resulting number that minimises it, are in agreement with what was observed from each repetition's plots in Figure 7.8.

The 'tied' covariance type was found more suitable for this application during the analysis on the individual substations that was discussed earlier in this chapter, therefore it was selected for the case of multiple substations as well. Having obtained the optimal number of components, the GMM with tied covariance and 63 components was used to assign each selected substation day to one of these profiles. Then, the model learned on the selected days, was used to label each of the remaining substation days of the initial dataset. The 63 representative load current profiles that were obtained from the application of Gaussian Mixture Models to data from 75 substations are shown in Figure 7.10.

The next step was to investigate whether the methodology that was applied in individ-

Figure 7.10: 63 components of the GMM resulted from applying the method to load current data from 75 substations. For each subfigure the x and y axes show the hour of day and the load current respectively.

ual substations and was presented in the previous sections would work in this case as well. The predictive model used was again Gradient Boost and it was applied on data from each of the substations, using the same inputs as before. The labels corresponding to each substation were a subset of the 63 labels produced by the GMM and the numbers used to refer to them are common across all substations. As seen in Section 7.2, the prediction of days with phase overcurrent events gave good results in three substations (#1, #3 and #4), while the prediction of voltage swell event was good or relatively good for 11 substations (#2, #4, #5, #14, #18, #21, #23, #26, #29, #30 and #43). The results for the phase overcurrent and voltage swell event day prediction for the above substations were presented in Tables 7.3 and 7.4 respectively. Tables 7.6 and 7.7 show the results for the same substations but using the common labels across substations as was described in this section.

Table 7.6: Results for Common Labels Across Substations (Phase Overcurrent Events)

Substation number	**Accuracy**	**Precision (Event)**	**Recall (Event)**	**Precision (No Event)**	**Recall (No Event)**	**Results**
1	0.585	0.588	0.303	0.984	0.0189	bad
3	0.709	0.611	0.817	0.735	0.693	relatively good
4	0.742	0.728	0.756	0.699	0.778	good

Table 7.7: Results for Common Labels Across Substations (Voltage Swell Events)

Substation number	**Accuracy**	**Precision (Event)**	**Recall (Event)**	**Precision (No Event)**	**Recall (No Event)**	**Results**
2	0.822	0.714	0.903	0.837	0.813	good
4	0.663	0.661	0.683	0.791	0.510	relatively good
14	0.739	0.718	0.866	0.960	0.374	relatively good
18	0.728	0.730	0.671	0.672	0.795	relatively good
21	0.674	0.647	0.721	0.788	0.557	relatively good
23	0.672	0.573	0.722	0.496	0.776	relatively good
26	0.855	0.846	0.859	0.521	0.966	good
29	0.640	0.595	0.650	0.251	0.588	bad
30	0.705	0.646	0.761	0.698	0.710	relatively good
43	0.906	0.702	0.915	0.218	0.987	bad

A comparison of these results with those presented in Section 7.2 shows that for the case of phase overcurrent events, where all 3 substations had previously achieved good results, only 1 substation (#4) still gave good results (although slightly worse than

before). From the remaining 2 substations, 1 gave relatively good results (#3) and 1 gave bad results (#1) .

For the case of voltage swell events, where a higher number of substations had previously achieved good or relatively good results, the following observations can be made. First, substation #5 is not included in Table 7.7, as an error prevents the predictive model to run. For the remaining substations, 6 had previously achieved good prediction results while for 4 the results were relatively good. From the 6 substations with good results, the results stayed good in 2 of them (#2, #26) , while in 3 substations the results were relatively good (#14, #21, #30) and in 1 they were bad (#43) . From the 4 substations with relatively good results, the results remained relatively good in 3 (#4, #18, #23) , while the prediction results on the last substation were bad (#29) .

The fact that in 3 substations (from both tables), the results went from good or relatively good to bad when the load profiles were obtained across all substations could indicate overfitting for these three substations in the case where individual substations were considered. The comparison of the results for the cases where single and multiple substations were considered, also showed that there were substations where the results went from good to relatively good, or remained good or relatively good, but the values were slightly lower for some of the metrics. This small decrease in performance may be due to the fact that not all 24120 substation days were considered in order to obtain the representative load profiles and wrong labels might have been assigned to some days. To check this assumption, the label probabilities are investigated.

As explained in Chapter 4, in GMM clustering, all data points belong to all clusters with a certain probability. For each of these data points, the highest probability is found and the data point is assigned to the cluster that corresponds to that probability. The label probabilities for the 24120 substation days in the dataset are shown in Figure 7.11.

It is evident from the histogram that the majority of labels were assigned with a very high probability, which in most cases was above 90%. However, there are some cases, where the probability used for label assignment was below 50%. This could be explained

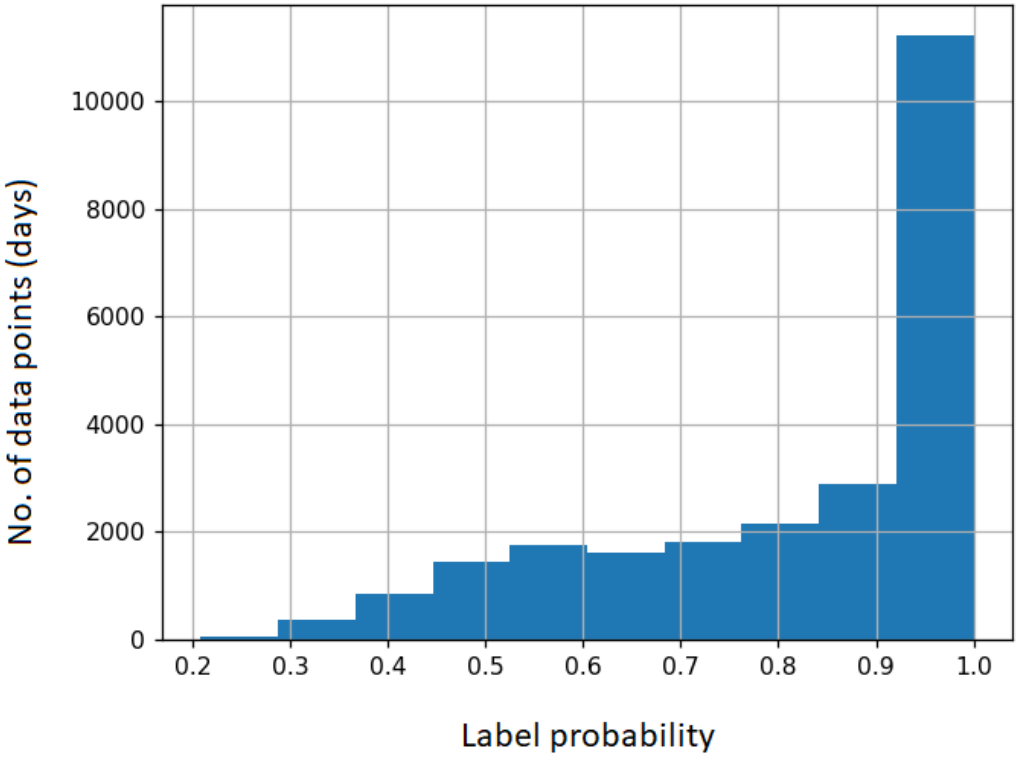


Figure 7.11: Histogram of label probabilities for the 24120 substation days.

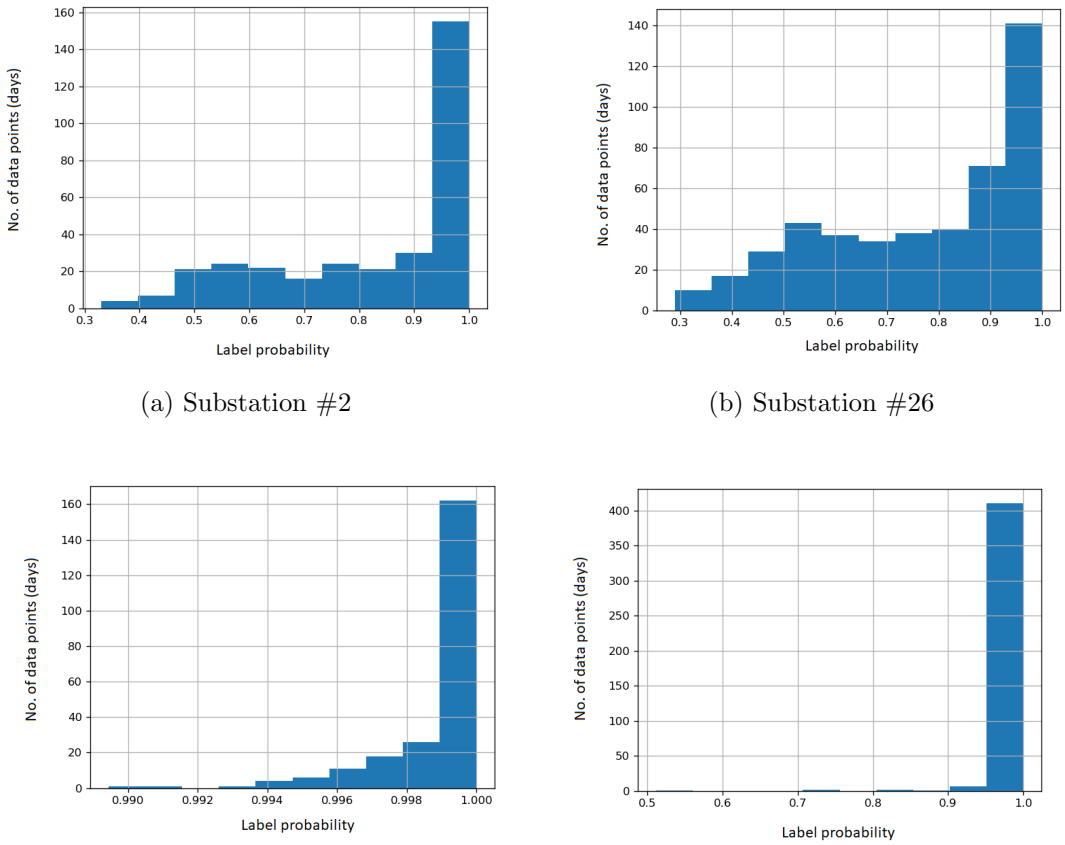
by the fact that only 10% of the substation days were fed into the GMM in order to obtain the representative load profiles and there might be some less common profiles that have been missed.

To check how this can affect the performance of the models, the predictive model is applied again on the substation data but, in this case, only the substation days with label probabilities higher than 80% are considered and the results were the following:

- In 2 substations the results were better when only the days with high label probability were considered (substations #2 and #26)
- For the majority of substations the results were similar to the case where substation days with all label probability values were considered
- The results remained bad for the 3 substations with bad results discussed above
- Worse results were obtained for one substation (substation #30 , where the results were worse than before but still relatively good)

The histograms in Figure 7.12 show the label probabilities of all days for the two substations, for which the performance was better when only the higher label probabilities were considered (shown at 7.12a and 7.12b), and for two of the substations, for which the prediction results were very similar in the two cases (shown at 7.12c and 7.12d). It is clear from the histograms that the performance of the predictive models can be adversely affected when there are substation days, for which the label probabilities are low, as this is an indicator of wrongly assigned labels. Table 7.8 shows the results for substation #2 for the three cases considered:

- Case 1: Representative load profiles obtained from applying a GMM per substation and using the resulting labels for prediction
- Case 2: Representative load profiles obtained from applying a GMM on load current data from all available substations and using the labels (which were common across substations) to predict an *Event Day* per substation
- Case 3: Load profiles obtained in the same way as in Case 2 but only the days



(a) Substation #2

(b) Substation #26

(c) Substation #3

(d) Substation #21

Figure 7.12: Label probabilities for all substation days for four substations.

with label probabilities higher than 80% were used for prediction

The comparison of the results presented in the above table shows that, for certain substations, the prediction performance is higher when the labels have been assigned to each day with a high confidence. This means that the representative load profiles obtained by the application of a GMM on multiple substations and used as labels for a specific substation, can accurately describe the load behaviour of that substation.

7.5 Conclusion

The potential relationship between the changing load behaviour of secondary substations and power quality event occurrence was investigated in this chapter. The

Table 7.8: Comparison of Prediction Results on Substation #2 for Cases $1-3$

Labels used	**Accuracy**	**Precision (Event)**	**Recall (Event)**	**Precision (No Event)**	**Recall (No Event)**	**Results**
Case 1	0.833	0.717	0.890	0.932	0.801	good
Case 2	0.822	0.714	0.903	0.837	0.813	good
Case 3	0.861	0.747	0.899	0.945	0.841	good

methodology demonstrated here involved the application of Gaussian Mixture Models to identify a number of representative load profiles and then Gradient Boost was used to classify the substations into *Event Day* or *No Event Day* based the load profile of each day and the profile that it transitioned from. It was found that it is possible to predict an *Event Day* when there were enough (but not too many) event days available to train the predictive models. It was also found that t-SNE visualisation was very accurate for this type of application as the separation of data performed by t-SNE agreed with the partitioning specified by the GMM clustering. The good prediction results were achieved when the prevalence of event days (ratio of event days over total number of days) of certain power quality events was in the range 0.18 - 0.77. Outside this range, the results were generally bad, as in the case of lower prevalence (below 0.18) there were not enough *Event Day* examples and the model learned to classify almost all days as *No Event Days*, while in the case where most of the days had a recorded event (prevalence above 0.77) almost all days were classified as *Event Days* by the model. The prediction results presented in this chapter refer to phase overcurrent and voltage swell events only, as these were the only event types with an adequate number of events recorded. However, it is possible that this methodology would work with other event types as well, as they could also be driven by load behaviour. Two different ways of obtaining the representative load profiles using GMM were examined and compared. In both cases, where the profiles obtained for a single substation or were common across substations, the methodology gave satisfactory results for certain substations. The results of the first case and the comparison with those of the second showed that looking into the individual substations separately tends to ovefit but it may also be preferable for certain substations with unusual loading that cannot be accurately represented by the profiles obtained when GMM were applied to multiple

substations. The case of multiple substations showed that when the labels (profiles) have been assigned to the substation with a high degree of certainty, Gradient Boost can predict the occurrence of certain power quality events, while certain substations for which the predictive models tend to overfit could also be identified. Being able to use load profiles from many substations and then use them to predict an *Event Day* could be beneficial as the learning could be subsequently applied to unmonitored substations as well. This would work better if substations of similar characteristics and loading conditions were grouped together and fed to a GMM to obtain the representative load profiles for that specific group and then apply the knowledge regarding the profiles that are more likely to be associated with a power quality event to other substations with similar characteristics and loading to that group.

Chapter 8

Conclusion

The topic of Fault Anticipation in Distribution Networks was addressed in this thesis. Rather than focusing on the prediction of specific types of electrical faults using the relevant operational data, this thesis considered 'Fault Anticipation' in a wider sense, where the prediction of faults or power quality disturbances is explored through the utilisation of existing data coming from power systems or other sources (e.g. weather data), while taking into account the changing nature of distribution networks and the current and anticipated data related challenges.

Power systems are currently undergoing a transformation process towards the power networks of the future, which are known as smart grids, with large part of this transformation occurring at the distribution level. Contrary to the traditional power systems, where the distribution networks were passive and their main role was to accommodate the uni-directional power flow from substations to consumers, future distribution networks will be characterised by bi-directional power and information flow and they will heavily rely on monitoring systems and data utilisation to be able to accommodate the large scale adoption of LCTs as well as the anticipated increase in demand. In order to achieve a successful transition to the future distribution network and to fully utilise the anticipated amounts of data generated in the future, DNOs will need to have a suitable data management and analysis strategy. Although the current state of monitoring at the distribution level is far less advanced in practice compared to what is described

above, DNOs should start building their future data strategy now. Currently, little to no monitoring is usually available for large parts of UK distribution network, while in other parts of it is common for the DNOs to have monitoring equipment in place and never look into the recorded data. Acknowledging the challenge of limited data availability and the rapidly changing operating nature of distribution networks, the purpose of this thesis was to provide DNOs with recommendations and examples that could help with the development of this strategy. To this end, a general data analysis methodology, which involves repurposing of existing data in order to extract additional value, was proposed and demonstrated with a series of short and more detailed case studies.

The short case studies addressed in this thesis involved PV identification, phase imbalance and detection of unusual network behaviour. These three topics were identified during the exploration and analysis of available data and served as examples of how the proposed methodology can be used to identify potential issues or useful information that may be hidden in the data. The more detailed case studies, which form the main part of the thesis, have focused on the following two areas: (i) Prediction of weather-related faults on minimally observed distribution networks and (ii) Impact of substation loading on the occurrence of power quality disturbances. The research in both areas had the common goal to utilise machine learning in order to develop a methodology towards the prediction of distribution network power quality events or faults in the absence of extensive monitoring. Both of these research topics stemmed from domain knowledge and the application of the proposed methodology on these areas aimed to translate the tacit knowledge, which suggested that certain factors lead to the occurrence of certain events, to a formal model that confirms these conjectures.

Before the presentation of the above case studies, the current and anticipated data related challenges encountered in practice and from related literature were reviewed and recommendations for a more efficient data management strategy were made. It was pointed out that, although long term decisions must be made to be able to fully utilise the anticipated increase in data, it is equally important for DNOs to improve

their data management now so that they can maximise the value of existing data and then consider how to adapt their developed strategy to accommodate the data requirements of the future network. The proposed data analysis methodology consists of three stages: Characterising Network Behaviour, Anomaly Detection and Prediction of Faults or Disturbances. The purpose of each stage of the analysis was discussed in Chapter 4, where selected examples of data analysis and machine learning techniques that can be used for each of these stages were also explained. The methodology was applied to real data and demonstrated through three short and two more detailed case studies.

8.1 Outcomes of Research

The main observations and results obtained from the short and more detailed case studies presented in this thesis are summarised in this section.

Unusual Network Operation

The three short case studies presented in Chapter 5 involved the application of various techniques in order to detect unusual operation of the network. In the first case study, a different operating condition of phase A compared to that of phases B and C was identified after a comparison of the load currents measured at 6 monitoring locations. The nature of the observed behaviour and the fact that it coincided with increased solar irradiance, led to the assumption that it was caused by solar PV operation. Next, in the second case study, the use of a ternary plot to identify the presence of unbalance was demonstrated. Further examination identified a potential relation between ambient temperature and unbalance, as unbalance seemed to increase with higher temperature. The detection of both PV generation and unbalance are of interest to the DNOs as being able to detect unmetered PV generation on their networks would allow them to be prepared for potential voltage and power quality issues, while in the case of unbalance, measures to mitigate the impact of unbalanced load such as losses could be taken. The third case study of Chapter 5 demonstrated how a dimensionality reduction technique such as t-SNE could be applied on high-dimensional distribution network data in order

to identify unusual behaviour of the network. Looking at the results, a number of observations were made: Even with 15-minute data, it was possible to detect signs of unusual behaviour within the data. Spatial and seasonal variation was observed as the behaviour of two neighbouring substations was found to be different than that of other substations and that behaviour was very different at a specific time of the year. The purpose of the short case studies was to provide illustrative examples of how the proposed methodology would be used by a DNO. However, as the main topic of this thesis was the prediction of faults, more emphasis was placed on the more detailed case studies of Chapters 6 and 7.

Weather-Related Fault Prediction Using Only Weather Observations

The work presented in Chapter 6 addressed the topic of fault prediction on networks with minimal monitoring. After a brief discussion on the fault history of the distribution network considered and the available data, the proposed methodology towards the prediction of weather related faults using only weather data and its application on a real distribution network were presented. The results were presented with three case studies, where the performance of different classification methods on datasets with varying input variables was compared. The first two case studies involved all weather faults occurred at the HV and LV levels. Linear Discriminant Analysis was the best performing method for weather-related fault prediction at the HV level, with an accuracy of 79.2% for both *Fault / No Fault* classification and classification based on the fault cause. For the LV level, Gradient Boost performed better in *Fault / No Fault* classification for weather-related faults with and accuracy of 82%. The above results showed that it is possible to predict the occurrence of a weather-related fault at a specific part of the network using only weather variables. The final case study on a subset of unknown cause LV faults showed that there could be a higher number of faults caused by weather conditions than those registered as such, as prediction results on the unknown cause dataset were also good, although fewer weather variables were considered for that case. This could also be used to retrospectively analyse persistent faults with unknown causes and potentially attribute them to certain weather conditions.

The contribution and novelty of this work is a methodology for finding the functional relation between fault occurrence and environmental conditions. A practical use case stemming from this methodology would be using the model with a longer term weather forecast to understand which parts of the network were at risk of fault under the forecasted weather conditions. This would assist in the refinement of spares budgets and strategic positioning of maintenance staff although at shorter timescales. With this use case in mind, the benefits for network operators could be further enhanced by moving the methodology towards a probabilistic framework which would in turn accommodate uncertainties in forecasts and measurement errors to provide probability of fault rather than just prediction. As distribution network operators face increasingly diverse challenges on their ageing infrastructure, such an approach would allow them to act on predictions according to their attitude to risk which in turn could be informed by asset health and criticality indices.

Power Quality Event Prediction Using Load Behaviour

In Chapter 7, the potential relationship between the changing load behaviour of secondary substations and power quality event occurrence was investigated. The methodology demonstrated here involved the application of Gaussian Mixture Models to identify a number of representative load profiles and then Gradient Boost was used to classify the substations into *Event Day* or *No Event Day* based the load profile of each day and the profile that it transitioned from. It was found that it is possible to predict an *Event Day* when there were enough (but not too many) event days available to train the predictive models. It was also found that t-SNE visualisation was very accurate for this type of application as the separation of data performed by t-SNE agreed with the partitioning specified by the GMM clustering. The good prediction results were achieved when the prevalence of event days (ratio of event days over total number of days) of certain power quality events was in the range 0.18 - 0.77. The prediction results presented in this chapter refer to phase overcurrent and voltage swell events only, as these were the only event types with an adequate number of events recorded. However, it is possible that this methodology would work with other event types as

well, as they could also be driven by load behaviour. Two different ways of obtaining the representative load profiles using GMM were examined and compared. In both cases, where the profiles obtained for a single substation or were common across substations, the methodology gave satisfactory results for certain substations. The results of the first case and the comparison with those of the second showed that looking into the individual substations separately tends to ovefit but it may also be preferable for certain substations with unusual loading that cannot be accurately represented by the profiles obtained when GMM were applied to multiple substations. The case of multiple substations showed that when the labels (profiles) have been assigned to the substation with a high degree of certainty, Gradient Boost can predict the occurrence of certain power quality events, while certain substations for which the predictive models tend to overfit could also be identified. Being able to use load profiles from many substations and then use them to predict an *Event Day* could be beneficial as the learning could be subsequently applied to unmonitored substations as well. This would work better if substations of similar characteristics and loading conditions were grouped together and fed to a GMM to obtain the representative load profiles for that specific group and then apply the knowledge regarding the profiles that are more likely to be associated with a power quality event to other substations with similar characteristics and loading to that group.

As seen in the chapters addressing the weather-related fault prediction and the impact of loading on the occurrence of power quality events, the developed models showed a good performance that could be used to inform decisions. It is proposed that these models use weather or load forecasts as inputs in order to predict the occurrence of a weather-related fault or a power quality event, such as voltage swell, respectively. Although a longer term forecast could provide an insight of the events that would be expected under the forecasted conditions, it is suggested that the predictions be updated using day ahead forecasts, as they could more accurately inform DNOs in order to decide on the actions that need to be taken.

8.2 Future Work

There are multiple directions that this work can be extended and applied, including replicating similar analysis for different distribution networks (other countries, DNOs) or extending this work for other data, and testing the overall methodology proposed. The data management challenges in DNOs were discussed in Chapter 3, which pointed out the main challenges and gave recommendations. Future work could involve further investigation of this topic by examining how DNOs could use these recommendations to develop an efficient data management strategy. Building on the Energy Data Taskforce work [71], which presented a vision of the future energy sector, where the full utilisation of data enables the transformation to a decarbonised and decentralised system, future work could take the recommendations made by the Energy Data Taskforce to the DNO level, investigate in detail their current status and identify potential barriers that prevent the DNOs from adopting the principles outlined in that work. This would help DNOs translate a set of recommendations to specific actions, which might be different for different DNOs.

In addition, future work could build on the work presented in Chapters 6 and 7. Indicatively, future work could involve the development of a weather fault prediction system utilising the methodology proposed and demonstrated in Chapter 6. Rather than giving a firm prediction result, this system could provide the user with the probability that a weather related fault might occur at a specific area of the network, given the forecasted weather conditions for that area. Regarding the work presented in Chapter 7, where a relation between substation loading and power quality event occurrence was established, a future direction could be the application of the demonstrated methodology on network faults, in order to investigate the impact of changing load on the occurrence of actual faults. Finally, the work presented in Chapter 7 could be extended to develop models that predict the occurrence of power quality events or faults at groups of substations with similar characteristics. These models, which will be obtained from monitored substations could then used to predict an event occurrence at an unmonitored substation that belongs to the same group (i.e. it has the same characteristics as

the substations used to develop the predictive model). Transfer Learning [96], which is an area of Machine Learning focusing on how to transfer knowledge gained from solving a particular problem to another could be used to apply the predictive models developed on monitored substations to others, where no monitoring is available.

References

[1] Stratfor Worldview, "Smart grids will revolutionize power distribution," https://worldview.stratfor.com/article/smart-grids-will-revolutionize-power-distribution, July 2014, accessed: 08-01-2020.

[2] Energy Networks Association, "Open Networks Project – DSO Definition and R&R," https://www.energynetworks.org/assets/files/electricity/futures/Open_Networks/ON-WS3-DSO%20Definition%20(updated)%20-%20published%20v1.pdf, June 2018.

[3] M. Zipf and D. Möst, "Cooperation of TSO and DSO to provide ancillary services," in *2016 13th International Conference on the European Energy Market (EEM)*. IEEE, 2016, pp. 1–6.

[4] Energy Networks Association, "Open Networks Project — DSO Transition: Roadmap to 2030," https://www.energynetworks.org/assets/files/electricity/futures/Open_Networks/DSO%20Roadmap%20v6.0.pdf.

[5] B. D. Russell, C. L. Benner, R. M. Cheney, C. F. Wallis, T. L. Anthony, and W. E. Muston, "Reliability improvement of distribution feeders through real-time, intelligent monitoring," in *2009 IEEE Power & Energy Society General Meeting*. IEEE, 2009, pp. 1–8.

[6] K. Muthu-Manivannan, C. L. Benner, P. Xu, and B. D. Russell, "Electrical power system event detection and anticipation," Jun. 4 2013, US Patent 8,457,910.

[7] J. A. Wischkaemper, C. L. Benner, B. D. Russell, and K. M. Manivannan, "Waveform analytics-based improvements in situational awareness, feeder visibility, and operational efficiency," in *2014 IEEE PES T&D Conference and Exposition*. IEEE, 2014, pp. 1–5.

[8] J. A. Wischkaemper, C. L. Benner, B. D. Russell, and K. Manivannan, "Application of waveform analytics for improved situational awareness of electric distribution feeders," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 2041–2049, 2015.

[9] K. Wong, H. Ryan, and J. Tindle, "Power system fault prediction using artificial neural networks," in *Neural Information Processing, International Conference on*. Springer, 1996, pp. 1181–1186.

[10] S. Zhang, Y. Wang, M. Liu, and Z. Bao, "Data-based line trip fault prediction in power systems using LSTM networks and SVM," *IEEE Access*, vol. 6, pp. 7675–7686, 2017.

[11] X. Wang, S. D. McArthur, S. M. Strachan, J. D. Kirkwood, and B. Paisley, "A data analytic approach to automatic fault diagnosis and prognosis for distribution automation," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6265–6273, 2017.

[12] S. M. Strachan, S. D. McArthur, B. Stephen, J. R. McDonald, and A. Campbell, "Providing decision support for the condition-based maintenance of circuit breakers through data mining of trip coil current signatures," *IEEE Transactions on Power Delivery*, vol. 22, no. 1, pp. 178–186, 2007.

[13] M. Kezunovic, Z. Ren, G. Latisko, D. R. Sevcik, J. S. Lucey, W. E. Cook, and E. A. Koch, "Automated monitoring and analysis of circuit breaker operation," *IEEE Transactions on Power Delivery*, vol. 20, no. 3, pp. 1910–1918, 2005.

[14] H. Hoidalen and M. Runde, "Continuous monitoring of circuit breakers using vibration analysis," *IEEE Transactions on Power Delivery*, vol. 20, no. 4, pp. 2458–2465, 2005.

[15] P. Gross, A. Boulanger, M. Arias, D. L. Waltz, P. M. Long, C. Lawson, R. Anderson, M. Koenig, M. Mastrocinque, W. Fairechio *et al.*, "Predicting electricity distribution feeder failures using machine learning susceptibility analysis," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 2, 2006, p. 1705.

[16] A. Asheibi, D. Stirling, and D. Sutanto, "Analyzing harmonic monitoring data using supervised and unsupervised learning," *IEEE Transactions on Power Delivery*, vol. 24, no. 1, pp. 293–301, 2009.

[17] T. Gu, P. Kadurek, J. Cobben, and A. Endhoven, "Power quality data evaluation in distribution networks based on data mining techniques," in *Environment and Electrical Engineering (EEEIC), 2013 12th International Conference on.* IEEE, 2013, pp. 58–63.

[18] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa, and T. Gozel, "Representative residential LV feeders: A case study for the North West of England," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 348–360, 2016.

[19] A. E. Lazzaretti, V. H. Ferreira, H. V. Neto, L. F. Toledo, and C. L. Pinto, "A new approach for event classification and novelty detection in power distribution networks," in *Proceedings of the IEEE PES General Meeting.* Citeseer, 2013.

[20] A. E. Lazzaretti, D. M. J. Tax, H. V. Neto, and V. H. Ferreira, "Novelty detection and multi-class classification in power distribution voltage waveforms," *Expert Systems with Applications*, vol. 45, pp. 322–330, 2016.

[21] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window PCA," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2537–2548, 2016.

[22] Y. Liu, L. Zhan, Y. Zhang, P. N. Markham, D. Zhou, J. Guo, Y. Lei, G. Kou, W. Yao, J. Chai *et al.*, "Wide-area-measurement system development at the distribution level: An FNET/GridEye example," *IEEE Transactions on Power Delivery*, vol. 31, no. 2, pp. 721–731, 2016.

[23] A. Kenward and U. Raja, "Blackout: Extreme weather, climate change and power outages," *Climate central*, vol. 10, 2014.

[24] Northern Powergrid, "Adapting to climate change," June 2015.

[25] J. M. Murphy, D. Sexton, G. Jenkins, B. Booth, C. Brown, R. Clark, M. Collins, G. Harris, E. Kendon, R. Betts *et al.*, "UK climate projections science report: climate change projections," 2009.

[26] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electric Power Systems Research*, vol. 127, pp. 259–270, 2015.

[27] M. Panteli, C. Pickering, S. Wilkinson, R. Dawson, and P. Mancarella, "Power system resilience to extreme weather: Fragility modelling, probabilistic impact assessment, and adaptation measures," *IEEE Transactions on Power Systems*, vol. 32, pp. 3747–3757, 2017.

[28] K. Murray and K. Bell, "Wind related faults on the GB transmission network," in *2014 International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2014*, 2014.

[29] G. Li, P. Zhang, P. B. Luh, W. Li, Z. Bie, C. Serna, and Z. Zhao, "Risk analysis for distribution systems in the northeast US under wind storms," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 889–898, 2014.

[30] B. Matic-Cuka and M. Kezunovic, "Improving smart grid operation with new hierarchically coordinated protection approach," 2012.

[31] M. Kezunovic, P. Chen, A. Esmaeilian, and M. Tasdighi, "Hierarchically coordinated protection: An integrated concept of corrective, predictive, and inherently adaptive protection," in *Proceeding 5th International Scientific and Technical Conference*, 2015.

[32] A. Dagnino, K. Smiley, and L. Ramachandran, "Forecasting fault events in power distribution grids using machine learning." in *SEKE*, 2012, pp. 458–463.

[33] P.-C. Chen, T. Dokic, and M. Kezunovic, "The use of big data for outage management in distribution systems," in *International Conference on Electricity Distribution (CIRED) Workshop*, 2014.

[34] P.-C. Chen, T. Dokic, N. Stokes, D. W. Goldberg, and M. Kezunovic, "Predicting weather-associated impacts in outage management utilizing the GIS framework," in *Innovative Smart Grid Technologies Latin America (ISGT LATAM), 2015 IEEE PES*. IEEE, 2015, pp. 417–422.

[35] P.-C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2827–2836, 2016.

[36] M. Tremblay, R. Pater, F. Zavoda, D. Valiquette, G. Simard, R. Daniel, M. Germain, and F. Bergeron, "Accurate fault-location technique based on distributed power-quality measurements," in *19th International Conference on Electricity Distribution*, 2007, pp. 21–24.

[37] R. Eskandarpour, A. Khodaei, and A. Arab, "Improving power grid resilience through predictive outage estimation," in *Power Symposium (NAPS), 2017 North American*. IEEE, 2017, pp. 1–5.

[38] D. Wang, R. J. Passonneau, M. Collins, and C. Rudin, "Modeling weather impact on a secondary electrical grid," *Procedia Computer Science*, vol. 32, pp. 631–638, 2014.

[39] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-time prediction of the duration of distribution system outages," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 773–781, 2019.

[40] T. Gu, J. Janssen, E. Tazelaar, and G. Popma, "Risk prediction in distribution networks based on the relation between weather and (underground) component failure," *CIRED-Open Access Proceedings Journal*, vol. 2017, no. 1, pp. 1442–1445, 2017.

[41] T. Dokic, M. Pavlovski, D. Gligorijevic, M. Kezunovic, and Z. Obradovic, "Spatially aware ensemble-based learning to predict weather-related outages in transmission," in *The Hawaii International Conference on System Sciences–HICSS*, 2019.

[42] Y. Zhou, A. Pahwa, and S.-S. Yang, "Modeling weather-related failures of overhead distribution lines," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1683–1690, 2006.

[43] A. I. Sarwat, M. Amini, A. Domijan, A. Damnjanovic, and F. Kaleem, "Weather-based interruption prediction in the smart grid utilizing chronological data," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 2, pp. 308–315, 2016.

[44] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templates —- Part I: Substation clustering and classification," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3036–3044, 2014.

[45] B. Akperi and P. Matthews, "Analysis of clustering techniques on load profiles for electrical distribution," in *2014 International Conference on Power System Technology*. IEEE, 2014, pp. 1142–1149.

[46] E. C. Bobric, G. Cartina, and G. Grigoras, "Clustering techniques in load profile analysis for distribution stations," *Advances in Electrical and Computer Engineering*, vol. 9, no. 1, pp. 63–66, 2009.

[47] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, 2012.

[48] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[49] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933–940, 2006.

[50] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[51] O. Y. Al-Jarrah, Y. Al-Hammadi, P. D. Yoo, and S. Muhaidat, "Multi-layered clustering for power consumption profiling in smart grids," *IEEE Access*, vol. 5, pp. 18 459–18 468, 2017.

[52] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120–1128, 2007.

[53] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911–918, 2014.

[54] R. Singh, B. C. Pal, and R. A. Jabr, "Statistical representation of distribution system loads using gaussian mixture model," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 29–37, 2009.

[55] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 88–96, 2013.

[56] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and markov models," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1561–1569, 2013.

[57] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[58] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency–domain load pattern data," *International Journal of*

Electrical Power & Energy Systems, vol. 28, no. 1, pp. 13–20, 2006.

[59] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied Energy*, vol. 141, pp. 190–199, 2015.

[60] S.-l. Yang, C. Shen *et al.*, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, 2013.

[61] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2437–2447, 2016.

[62] B. Auder, J. Cugliari, Y. Goude, and J.-M. Poggi, "Scalable clustering of individual electrical curves for profiling and bottom-up forecasting," *Energies*, vol. 11, no. 7, p. 1893, 2018.

[63] M. Lees and EA Technology Limited, "Customer-led network revolution – enhanced network monitoring report," December 2014.

[64] IET, "Electricity networks – handling a shock to the system," https://www.theiet.org/media/2785/elec-shock-tech.pdf, 2013.

[65] S. Sagiroglu, R. Terzi, Y. Canbay, and I. Colak, "Big data issues in smart grid systems," in *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*. IEEE, 2016, pp. 1007–1012.

[66] Committee on Climate Change, "UK housing: Fit for the future?" https://www.theccc.org.uk/wp-content/uploads/2019/02/UK-housing-Fit-for-the-future-CCC-2019.pdf, February 2019.

[67] F. C. Trindade, L. F. Ochoa, and W. Freitas, "Data analytics in smart distribution networks: Applications and challenges," in *2016 IEEE Innovative Smart Grid Technologies-Asia (ISGT-Asia)*. IEEE, 2016, pp. 574–579.

[68] R. Arghandeh and Y. Zhou, *Big data application in power systems.* Elsevier, 2017.

[69] IET, "An overview of Britain's changing energy sector," https://www.theiet.org/media/2307/energy-white.pdf.

[70] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 425–436, 2015.

[71] L. Sandys and Energy Data Taskforce, "A strategy for a modern digitalised energy system," https://es.catapult.org.uk/wp-content/uploads/2019/06/Catapult-Energy-Data-Taskforce-Report-A4-v4AW-Digital.pdf, June 2019.

[72] IEC61970, "Energy Management System Application Program Interface (EMS-API) part 301: Common Information Model (CIM) base," *Geneva, Switzerland: IEC*, 2003.

[73] IEC61968, "Aplication Integration at Electric Utilities – System Interfaces for Distribution Management – part 11: Common Information Model (CIM) Extensions for Distribution," 2013.

[74] D. Pyle, *Data preparation for data mining.* Morgan Kaufmann, 1999.

[75] J. W. Tukey, *Exploratory data analysis.* Reading, Mass., 1977, vol. 2.

[76] G. J. McLachlan and D. Peel, *Finite mixture models.* John Wiley & Sons, 2004.

[77] C. M. Bishop, *Pattern recognition and machine learning.* Springer, 2006.

[78] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[79] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[80] L. v. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[81] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[82] "Scikit learn – classifier comparison," http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html, accessed: 17-03-2020.

[83] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[84] T. M. Mitchell, *Machine learning*. Burr Ridge, IL: McGraw Hill, 1997.

[85] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[86] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[87] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[88] D. J. MacKay, "Probable networks and plausible predictions -— a review of practical bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–505, 1995.

[89] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[90] J. Aitchison, *The statistical analysis of compositional data*, ser. Monographs on statistics and applied probability. Chapman and Hall, 1986. [Online]. Available: https://books.google.co.uk/books?id=RHKmAAAAIAAJ

[91] Power Standards Lab, "PQube installation & user's manual," 2008 - 2012.

[92] C. Francisco, *Harmonics, power systems, and smart grids*. CRC Press, 2015.

[93] Electricity North West, "Fault restoration performance on HV rings as part of the capacity to customers project," 2013, http://www.enwl.co.uk/docs/default-source/c2c-key-documents/white-paper—fault-performance-of-hv-rings.pdf?sfvrsn=10.

[94] Met Office Website, "Spring 2013 weather summary," http://www.metoffice.gov.uk/climate/uk/summaries/2013/spring.

[95] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[96] S. J. Pan and Q. Yang, "A survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.